Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, Alexis Nasr

Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde

Ahmed Hamdi[1] Rahma Boujelbane[1,2] Nizar Habash[3] Alexis Nasr[1]
(1) Laboratoire d'Informatique Fondamentale de Marseille- CNRS - UMR 7279 Université Aix-Marseille
(2) Multimedia, InfoRmation Systems and Advanced Computing Laboratory, Sfax 3021, TUNISIE.
(3) Center for Computational Learning Systems Columbia University New York, NY 10115, USA
{ahmed.hamdi,rahma.boujelbane,alexis.nasr}@lif.univ-mrs.fr
habash@ccls.columbia.edu

RÉSUMÉ

Le développement d'outils de TAL pour les dialectes de l'arabe se heurte à l'absence de ressources pour ces derniers. Comme conséquence d'une situation de diglossie, il existe une variante de l'arabe, l'arabe moderne standard, pour laquelle de nombreuses ressources ont été développées et ont permis de construire des outils de traitement automatique de la langue. Etant donné la proximité des dialectes de l'arabe, le tunisien dans notre cas, avec l'arabe moderne standard, une voie consiste à réaliser une traduction surfacique du dialecte vers l'arabe moderne standard afin de pouvoir utiliser les outils existants pour l'arabe standard. Nous décrivons dans cet article une architecture pour une telle traduction et nous l'évaluons sur les verbes.

ABSTRACT

Translating verbs between MSA and arabic dialects through deep morphological analysis

The developpment of NLP tools for dialects faces the severe problem of lack of resources for such dialects. In the case of diglossia, as in arabic, a variant of arabic, Modern Standard Arabic, exists, for which many resources have been developped which can be used to build NLP tools. Taking advantage of the closeness of MSA and dialects, one way to solve the problem consist in performing a surfacic translation of the dialect into MSA in order to use the tools developped for MSA. We describe in this paper an achitecture for such a translation and we evaluate it on arabic verbs.

MOTS-CLÉS : dialectes, langues peu dotées, analyse morphologique, traitement automatique de l'arabe.

KEYWORDS: dialects, Arabic NLP, morphological analysis.

1 Introduction

Le monde arabophone connaît une situation de diglossie (Ferguson, 1959). Une forme d'arabe, l'arabe moderne standard (MSA) est partagée par tout le monde arabe, mais ne constitue la langue maternelle d'aucun arabophone. Le MSA est, en particulier, la langue de la presse écrite et parlée. D'autre part, il existe une grande variété de dialectes qui constituent les langues maternelles des arabophones. Les dialectes ne sont généralement pas écrits et ne possèdent par

conséquent pas de conventions orthographiques standard.

Cette situation particulière est problématique pour le traitement automatique des dialectes de l'arabe dans la mesure où les ressources pour ces langues sont quasiment inexistantes. En revanche, il existe des ressources importantes pour le MSA. L'idée que nous explorons dans cet article consiste à « traduire » un dialecte de l'arabe vers le MSA afin de pouvoir y appliquer des outils conçus pour le MSA. Le verbe traduire a été ici mis entre guillemets car l'objectif n'est pas d'obtenir une traduction parfaite mais une traduction de qualité suffisante pour appliquer des outils conçus pour le MSA. De façon plus précise, la traduction que nous proposons repose largement sur la morphologie et le lexique. C'est en effet à ces deux niveaux que se manifestent la majorité des différences entre les variétés de l'arabe. Le système proposé relève d'une architecture à transfert. Un mot en langue source est analysé sous la forme d'une racine, d'un schème [1] et de traits morphologiques. Un lexique bilingue permet alors de traduire la racine et le schème source vers une racine et un schème cible. Dans le cas, fréquent, où la racine est identique dans la langue source et la langue cible, la traduction se limite aux schèmes. La racine et le schème cible, ainsi que les traits morphologiques vont alors permettre de générer un ou plusieurs mots cibles. Nous nous limiterons, dans cet article, au traitement des verbes.

Une particularité de notre approche est de procéder à une analyse morphologique profonde, de manière à identifier la racine du mot cible alors que l'on aurait pu se contenter d'une analyse plus surfacique, sous la forme d'un lemme. La raison de ce choix est double. D'une part, la morphologie dérivationnelle de l'arabe est très régulière, l'identification de la racine peut être réalisée, de manière fiable et économique, à l'aide de règles. D'autre part, le fait de réaliser le transfert au niveau des racines permet de minimiser la taille du dictionnaire bilingue. On estime en effet à 7502 le nombre total de racines de l'arabe et 2903 racines fréquement utilisées (Altabbaa *et al.*, 2010), ce qui permet de définir une borne supérieure de notre dictionnaire. D'autre part, le système que nous proposons est bi-directionnel : tous les modules qui le composent sont réversibles, ce qui permet de réaliser la traduction depuis un dialecte vers le MSA et vice-versa [2].

Ce travail s'inscrit dans le contexte du traitement automatique des langues peu dotées, tel que les travaux de (Seng, 2010) sur le khmer et le laotien, ou les travaux de (Abdillahi *et al.*, 2006) sur le somali. Cependant, comme nous l'avons mentionné ci-dessus, la situation de l'arabe est particulière dans la mesure où les différentes variétés de l'arabe entretiennent une relation privilégiée avec le MSA pour lequel nous disposons de ressources importantes. En ce sens, notre travail se rapproche des travaux de (Scherrer *et al.*, 2009) sur les dialectes suisses allemands. L'auteur propose un système de traduction depuis l'allemand vers différents dialectes. Ce système repose sur une analyse syntaxique de l'allemand et c'est à l'issue de l'analyse syntaxique qu'un mécanisme de transfert permet de générer une traduction en dialecte. Notre approche se distingue de ces travaux par deux aspects importants. D'une part, le transfert dans notre cas est réalisé au niveau morphologique. Ce choix repose, comme nous l'avons vu, sur une hypothèse théorique (le niveau morphologique est un niveau de transfert acceptable dans notre cas) mais aussi sur une considération pratique qui est que l'on ne dispose pas d'un système d'analyse syntaxique pour le tunisien. Le second aspect qui distingue notre travail de (Scherrer *et al.*, 2009) est que notre système est bi-directionnel, il permet aussi bien de traduire du tunisien vers le MSA que l'inverse. Plus proche de nous linguistiquement, (Shaalan *et al.*, 2007) décrit un système de transfert de

1. Rappelons que l'arabe est une langue gabaritique. Les mots pleins de l'arabe peuvent être analysés sous la forme d'un gabarit ou schème et d'une racine.

2. La traduction du MSA vers un dialecte peut être intéressante dans une application de transcription automatique de la parole : on traduit en dialecte un corpus MSA afin de construire un modèle de langage pour le dialecte.

l'égyptien vers le MSA. Dans ce cas, le transfert est effectué au niveau des lemmes alors que nous l'effectuons au niveau des racines pour des raisons déjà évoquées ci-dessus.

La structure de l'article est la suivante : nous commencerons, section 2, par une très brève description de la morphologie verbale de l'arabe. La section 3 se penche sur la morphologie verbale du tunisien, en mettant en avant les aspects qui la distinguent de la morphologie verbale du MSA. La section 4 décrit l'outil MAGEAD dont nous nous sommes servis pour l'analyse et la génération morphologique. Nous décrivons ensuite, dans la section 5 notre lexique. Une évaluation du système est décrite en section 6 et la section 7 clôt l'article.

2 Morphologie verbale de l'arabe

Le système morphologique verbal de l'arabe est complexe : il met en jeu des phénomènes d'agglutination, de flexion et de dérivation. En revanche, il est très régulier, ce qui permet de le décrire de manière fiable et économique à l'aide de règles. L'objectif de cette section est de décrire brièvement les différents aspects de la morphologie verbale de l'arabe, en particulier les notions de clitiques, d'affixes, de lemmes, de racines et de schèmes. Ces notions nous permettront, en 3, de décrire de manière précise les différences entre la morphologie verbale du MSA et du tunisien et, en 4, d'introduire le système d'analyse et de génération morphologique que nous utilisons.

Dans la suite de cet article, nous présenterons nos exemples en alphabet arabe et sous une forme translitérée mise entre crochets. Pour cela, nous utilisons la translitération proposée par (Buckwalter, 2004).

2.1 Agglutination

La langue arabe est fortement agglutinante : des articles, des conjonctions, des prépositions, matérialisés par des **clitiques**, se rattachent aux formes fléchies. On distingue généralement les **proclitiques** qui se situent avant la forme fléchie et les **enclitiques** qui se situent après. Les clitiques sont optionnels et invariables (leur forme ne varie pas selon le verbe auquel ils se rattachent).

Le verbe arabe admet un seul enclitique, le pronom complément d'objet direct (PRN_D), qui varie en genre et en nombre et les proclitiques suivants présentés selon leurs positions, du plus éloigné au plus proche du verbe :

- QST : la particule d'interrogation أَ [>a] "*est-ce que*"
- CNJ : les conjonctions وَ [wa] "*et*" et فَ [fa] "*alors*"
- PRP : la préposition لِـ [li] "*pour*" et la particule d'accentuation لَـ [la].
- PRT : la particule de futur سَـ [sa] et les particules de négations لا [lA] et ما [mA]

La structure d'un verbe arabe peut être décrite par l'expression régulière suivante :

QST ? CNJ ? PRP ? PRT ? forme fléchie PRN_D ?

Illustrons cela sur le verbe أَسَتَكتُبُونَهَا [>asataktubuwnahA], qui se traduit en français par "*est-ce que vous l'écrirez*". Ce verbe est composé de deux proclitiques, l'article d'interrogation أَ [>a] et

la particule de futur سَـ [sa], une forme fléchie تَكتُبُونَ [taktubuwna] et un enclitique pronom d'objet direct هَا [hA].

L'opération qui consiste à séparer les clitiques du verbe est généralement appelée segmentation. Celle-ci pose des problèmes d'ambiguïté dans une perspective de traitement automatique. En effet, dans certains cas, plusieurs segmentations sont possibles, comme dans le cas du verbe وعده [wEdh] qui peut être décomposé en wEd+h "*il l'a promis*" ou bien comme w+Ed+h "*et il l'a compté*". L'ambiguïté est plus importante lorsque les diacritiques ne sont pas représentés, comme c'est généralement le cas dans les corpus arabes.

2.2 Flexion

La flexion verbale de l'arabe est très régulière. Elle est fondée sur la concaténation d'affixes aux lemmes verbaux. La détermination des affixes repose sur les valeurs des traits morphologiques suivants :

- Aspect : l'arabe distingue trois aspects : **le perfectif** utilisé quand l'action est accomplie. C'est l'aspect le plus simple d'un point de vue morphologique. Utilisé avec la troisième personne du singulier, il représente la forme canonique d'un verbe, à l'instar de l'infinitif en français. **L'imperfectif** indique que l'action est en train de se réaliser, sans être achevée. Il exprime le présent, et permet d'exprimer le passé et le futur à l'aide des particules. **L'impératif** indique l'injonction. Il ne peut être conjugué qu'à la deuxième personne.
- Mode : **l'indicatif** employé dans une proposition principale. **Le subjonctif** employé dans une proposition subordonnée. **Le jussif** ou l'apocopé exprime la négation, l'interdiction ou le conditionnel. Le mode s'applique uniquement à l'aspect imperfectif.
- Personne, genre et nombre du sujet : comme en français, on distingue trois personnes, deux genres, **le masculin** et **le féminin**. En revanche, l'arabe distingue trois valeurs pour le nombre **le singulier**, **le duel** et **le pluriel**.

Le tableau 1, décrit les affixes de la première personne selon le nombre, l'aspect et le mode du verbe. Le duel, l'impératif et le genre n'interviennent pas quand il s'agit de la première personne.

personne	nombre	Aspect	Mode	préfixe	suffixe	Exemple [katab]
1	singulier	perfectif	-	-	tu	katab**tu**
		imperfectif	indicatif	>	u	**>**aktub**u**
			subjonctif	>	a	**>**aktub**a**
			jussif	>	o	**>**aktub**o**
	pluriel	perfectif	-	-	nA	katab**nA**
		imperfectif	indicatif	n	u	**n**aktub**u**
			subjonctif	n	a	**n**aktub**a**
			jussif	n	o	**n**aktub**o**

TABLE 1: Affixes de flexion des verbes arabes pour la première personne

2.3 Racines et schèmes

Les lemmes verbaux arabes sont dérivés à patir d'une racine et d'un schème. La racine est une séquence de trois ou quatre lettres qui définit une notion abstraite. La racine كتب [ktb], par exemple, est associée à la notion d'écriture alors que la racine درس [drs] et liée à la notion d'étude. Un schème, appelé aussi gabarit ou patron, est une séquence composée de chiffres et de lettres qui définit le format du lemme. Le processus de génération d'un lemme consiste à remplacer chaque chiffre du schème par la lettre correspondante dans la racine. Reprenons l'exemple du lemme verbal كَتَبَ [katab], il est obtenu à partir de la racine ك ت ب ktb et le schème 1a2a3 en remplaçant, les chiffres 1, 2 et 3, par les lettres correspondantes de la racine.

Un schème est porteur d'un sens général, tel que le factitif, le nom prototypique de la personne qui effectue l'action, le résultat de l'action. . .le schème marque aussi la voix (on distinge l'actif et le passif sans agent) et l'aspect.

Le tableau 2 représente quelques schèmes des verbes arabes pour l'aspect perfectif ou imperfectif ainsi que leurs significations. Nous avons indiqué entre parenthèse le schème de la voix passive.

perfectif	imperfectif	signification
1a2a3 (1u2i3)	a12a3 (u12a3)	sens de base
1a22a3 (1u22i3)	u1a22i3 (u1a22a3)	causalité
1A2a3 (1uw2i3)	u1A2i3 (u1A2a3)	réciprocité implicite
ta1A2a3 (tu1uw2i3)	ata1A2a3 (uta1A2a3)	réciprocité explicite
1a23a4 (1u23i4)	u1a23i4 (u1a23a4)	sens de base
ta1a23a4 (tu1u23i4)	ata1a23i4 (uta1a23a4)	forme réfléchie de 1a23a4

TABLE 2: Exemples de schèmes verbaux arabes

3 Morphologie verbale du tunisien

Plusieurs travaux récents s'intéressent au dialecte tunisien : (Mejri *et al.*, 2009) a présenté la situation linguistique en Tunisie en décrivant les systèmes phonologiques, morphologiques et syntaxiques du tunisien. (Ouerhani, 2009) a étudié les phénomènes d'interférence entre la morphologie verbale du tunisien et celle de l'arabe standard d'une part, et la relation entre les verbes tunisiens et français (le cas de l'emprunt) d'autre part. Dans ce travail, nous nous intéressons tout comme (Ouerhani, 2009) à la morphologie verbale du dialecte tunisien mais contrairement à lui, qui ne s'interresse qu'à un échantillon de verbes, nous étudions tout le paradigme verbal tunsien. Ce dernier s'inspire fortement du MSA, on retrouve en effet les phénomènes d'agglutination de flexion et de dérivation décrits dans la section 2 mais avec quelques différences que nous décrivons ci-dessous.

3.1 Agglutination

Au niveau de l'agglutination, deux phénomènes distinguent le tunisien du MSA. D'une part certains clitiques MSA sont réalisés sous la forme de particules indépendantes en tunisien et

vice-versa. D'autre part, la forme de certains clitiques change. Ces phénomènes sont décrits plus en détails ci-dessous :

— le proclitique d'interrogation MSA أ [>a] "*est-ce que*" devient en tunisien l'enclitique ش [$] La forme verbale MSA أكتبتَ [>akatabta] "*est-ce que tu as écrit*", par exemple, se traduit en tunisien par كتبتش [ktibti$].
— la préposition لـ [li] "*pour*" et le proclitique du futur ne sont plus rattachés aux verbes. Tous les deux se traduisent par la particule indépendante باش [bA$] qui se situe avant le verbe : les formes لتكتُب [litaktub] "*pour que tu écrives*" et ستكتُب [sataktub] "*tu écriras*" sont exprimés en tunisien par باش تكتب [bA$ tiktib].
— le pronom complément d'objet indirect (PRN_I) qui est détaché du verbe en MSA se réalise sous la forme d'un enclitique en tunisien, par exemple les deux formes كَتَبتُ لكَ [katabtu laka] en MSA sont rattachées en tunisien كتبتلك [ktibtlik] "*je t'ai écrit*".

La structure d'un verbe tunisien peut être décrite par l'expression régulière suivante :

CNJ ? PRT ? forme fléchie PRN_D ? PRN_I ? (NEG|QST) ?

3.2 Flexion

De manière générale, la flexion des verbes tunisiens est plus pauvre que celle des verbes MSA. En particulier, le mode n'est plus marqué, les valeurs du nombre qui étaient au nombre de trois en MSA (singulier, duel et pluriel) sont réduits à deux (singulier et pluriel). Quant au genre, il n'est spécifié que lorsqu'il s'agit de la troisième personne du singulier. La liste des affixes sujet de la première personne sont représentés dans le tableau 3. Ce dernier peut être mis en regard du tableau 1.

personne	nombre	Aspect	préfixe	suffixe	Exemple : ktib "*écrire*"
1	singulier	perfectif	-	t	ktib**t**
		imperfectif	n	o	**n**iktib**o**
	pluriel	perfectif	-	nA	ktib**nA**
		imperfectif	n	uwA	**n**iktb**uwA**

TABLE 3: Affixes de flexions des verbes tunisiens pour la première personne

D'autre part, contrairement au MSA qui marque la voix dans le schème verbal, le tunisien marque la voix passive sous la forme du préfixe ت [t] [3]. La forme MSA passive كُتِبَ [kutiba] "*il est écrit*" devient en tunisien تِكتِب [tiktib].

3.3 Racines et schèmes

Hormis les emprunts, les lemmes verbaux tunisiens dérivent d'une racine et un schème, comme pour le MSA. Il y a en général correspondance bi-univoque entre un schème MSA et un schème tunisien sauf dans certains cas où un schème MSA peut correspondre à deux schèmes tunisiens

3. Nous aurions aussi pu définir le passif avec les schèmes, en ajoutant un /t/ au début de chaque schème de la voix active.

ou bien à aucun schème tunisien. La correspondance entre les schèmes MSA présentés dans la section 2 et les schèmes tunisiens est donnée dans le tableau 4.

perfectif		imperfectif	
schème_MSA	schème_TUN	schème_MSA	schème_TUN
1a2a3	12a3	a12a3	a12a3
1a22a3	1a22a3	u1a22i3	1a22a3
1A2a3	1A2a3	u1A2i3	1A2a3
ta1A2a3	t1A2a3	ata1A2a3	it1A2a3
1a23a4	1a23i4	u1a23i4	1a23i4
ta1a23a4	ta1a23i4	ata1a23i4	ta1a23i4

TABLE 4: Correspondance des schèmes MSA et tunisiens

4 Analyse et génération morphologiques

L'analyse et la génération morphologiques de notre système sont réalisées par l'outil MAGEAD (Habash et Rambow, 2006; Habash *et al.*, 2005). Ce dernier est un système à base de règles qui permet de décrire les systèmes morphologiques des différentes variétés de l'arabe (dialectes et MSA) et de les compiler sous la forme d'un transducteur fini.

Une des idées maîtresses qui sous-tendent le système MAGEAD est le partage des connaissances linguistiques communes à plusieurs variétés de l'arabe. En effet, comme nous l'avons vu ci-dessus, les variétés de l'arabe se distinguent par certains aspects lexicaux et morphologiques mais en partagent d'autres. L'architecture de MAGEAD permet de ne représenter qu'une fois ce qui est commun à plusieurs variétés de l'arabe.

MAGEAD effectue une analyse morphologique profonde. Partant d'une forme verbale ou nominale de l'arabe, il en fait l'analyse sous la forme d'une racine, d'une classe et de traits morphologiques. Ces derniers sont au nombre de 9 : `PER`, `GEN`, `NUM`, `ASP`, `VOICE`, `QST`, `CNJ`, `PRT`, `PRN`. Les cinq premiers traits définissent respectivement la personne, le genre, le nombre, l'aspect et la voix. Alors que les quatre derniers traits indiquent les clitiques (question, conjonction, particule et pronom d'objet direct). La combinaison de ces traits va permettre de sélectionner un schème, des affixes, des clitiques et de les combiner afin de produire une forme verbale.

MAGEAD distingue quatre niveaux de représentation. Nous les décrirons ci-dessous en nous appuyant sur un exemple, qui est la forme **ازدهرت** [Aizdaharat], "*elle a prospéré*".

— la représentation profonde.
A ce niveau de représentation, une forme est représentée, comme nous l'avons mentionné ci-dessus, sous la forme d'une racine, d'une classe, appelée MBC (pour *Morphologic Behavioural Class*) et de traits morphologiques. Ce niveau est commun à toutes les variantes de l'arabe.
A ce niveau, notre exemple est représenté sous la forme suivante :
`[ROOT:zhr] [MBC:verb-VIII] [POS:V] [PER:3] [GEN:f] [NUM:s] [ASP:p]`

— la représentation en morphèmes abstraits.

Les morphèmes abstraits sont des morphèmes qui pourront se réaliser différemment dans des variétés différentes de l'arabe.
Notre exemple est représenté à ce niveau de la façon suivante :
`[ROOT:zhr][PAT_PV:VIII][VOC_PV:VIII-act]+[SUBJSUF_PV:3FS]`
Les trois premiers morphèmes décrivent la racine, le schème (patron) et le vocalisme [4]. L'ensemble de ces trois morphèmes définissent un lemme. Le dernier morphème décrit un suffixe indiquant le genre, le nombre et la personne du verbe. Un tel suffixe pourra se réaliser différemment selon la variété d'arabe considérée.
Le passage du niveau profond au niveau morphologique abstrait est réalisé à l'aide des MBC. Ces derniers permettent d'associer des traits morphologiques à des morphèmes abstraits. Cette association est réalisée à l'aide de règles dont la partie gauche est constituée d'un ou plusieurs traits et la partie droite est constituée d'un morphème profond. C'est en particulier la règle suivante qui donnera naissance au morphème `[SUBJSUF_PV:3FS]` :
`[ASP:p][PER:3][GEN:f][NUM:s] -> [SUBJSUF_PV:3FS]`
Les MBC sont représentés sous la forme d'une hiérarchie, les MBC héritent de leurs MBC ancêtres un certain nombre de propriétés. C'est cette représentation hiérarchique qui permet de factoriser des règles communes à plusieurs MBC.

— la représentation en morphèmes concrets.
A ce niveau de représentation, les morphèmes abstraits sont réalisés sous la forme de morphèmes concrets. Notre exemple se représente maintenant de la façon suivante :
`<zhr, AV1tV2V3, iaa> + at`
le suffixe `+at` indique la personne, le genre et le nombre du sujet. Le triplet `<zhr, AV1tV2V3, iaa>` regroupe les trois composantes du lemme : la racine, le schème et le vocalisme. Ces trois composantes vont permettre de générer le lemme proprement dit. Le principe de génération relève de la morphologie non concaténative, elle consiste à remplacer les symboles 1,2 et 3 du schème par le permier, second et troisième symbole de la racine. Les symboles V sont quant à eux remplacés par les symboles qui consituent le vocalisme. Le résultat de cette opération est la chaîne `Aiztahra`. Cette opération est réalisée à l'aide d'un automate multibande, à l'image de (Kiraz, 2000).

— la représentation de surface.
Il s'agit de la représentation orthographique. Notre exemple se représente maintenant sous la forme `Aizdaharat`, qui est une translitération de la forme arabe ازدهرت. Le passage de la représentation en morphèmes concrets à la représentation de surface met en jeu deux types d'opérations. D'une part la concaténation des affixes et d'autre part des règles morphophonémiques qui vont, par exemple, provoquer le voisement du son /t/ pour donner le son /d/.

L'adaptation de MAGEAD à une nouvelle variété de l'arabe se décompose en trois étapes.

La première consiste à créer la nouvelle hiérarchie des MBC spécifiques au dialecte décrit. Dans notre cas, nous avons défini, pour chaque schème tunisien, un nouvel MBC dans la hiérarchie.

La deuxième étape consiste à définir de nouveaux morphèmes abstraits tels que, dans notre cas, l'enclitique de négation, ainsi que les morphèmes concrets leur correspondant. Dans le cas du tunisien la majorité des morphèmes concrets sont différents de ceux du MSA.

La troisième étape concerne les règles phonologiques et orthographiques propres au dialecte

4. MAGEAD ne manipule pas directement des schèmes, il les décompose en deux parties, d'une part une forme non diacritée du schème et d'autre part les diacritiques qui vont permettre de vocaliser cette forme afin d'obtenir un schème

décrit. Il existe en particulier une règle spécifique au tunisien qui remplace la troisième lettre de la racine, si cette dernière est défectueuse [5], ainsi que la voyelle qui la précède par la voyelle longue ا [A] lorsque le suffixe sujet commence par la voyelle fermée [u] ou [i] (ce qui est le cas pour la troisième personne du singulier féminin et la troisième personne du pluriel). Le verbe **مشى** [m$aY] conjugué à la troisième personne du singulier féminin donne **مشات** [m$At] alors qu'à la troisième personne du pluriel il donne **مشاوا** [m$AwA].

5 Lexique

Comme nous l'avons décrit dans l'introduction, notre lexique apparie des couples (racine, MBC) en MSA avec des couples (racine, MBC) en tunisien. Le lexique est composé de 1638 entrées. Il a été réalisé à partir du corpus de l'Arabic Tree Bank (ATB) (Maamouri *et al.*, 2004) qui est composé de 120 transcriptions d'émissions d'actualité en MSA diffusées par différentes chaînes arabes.

Ce corpus comporte 29911 occurences verbales. Afin d'extraire les lemmes et les racines de ces verbes, nous avons eu recours à l'analyseur morphologique ELIXIRFM (Smrž, 2007) qui permet, étant donné une forme fléchie en MSA, d'en extraire le lemme et la racine.

Chaque occurrence de lemme MSA a été ensuite traduite, en contexte, par un locuteur natif, en tunisien. A ce stade, les entrées du lexique sont composées, côté MSA d'un lemme et d'une racine et, côté tunisien, d'un lemme.

Nous avons alors associé à chaque entrée, du côté MSA, un MBC et pour chaque lemme, côté tunisien, un MBC et une racine. Comme nous l'avons décrit dans la section 4, lorsque le comportenement d'un verbe tunisien n'était pas décrit par un MBC MSA, un nouvel MBC a été créé.

En ce qui concerne les racines, dans 81,49 % des cas, nous avons identifié une racine arabe existante. Lorsqu'il n'existait pas de racine pour un lemme donné, nous avons eu recours à une méthode déductive pour en créer une nouvelle.

En effet, étant donné l'équation racine + schème = lemme, lorsque nous disposons d'un lemme et d'un schème, il est possible d'en déduire une racine. A l'aide de ce processus, nous avons défini une centaine de nouvelles racines spécifiques au tunisien.

Dans sa forme actuelle, le lexique est composé de 1638 entrées. Du côté tunisien l'ensemble des racines s'élève à 646 et du côté MSA à 1050.

L'ambiguïté est donc plus importante dans le sens tunisien → MSA que dans le sens MSA → tunisien. De manière plus précise, dans 587 cas, à un couple (racine, MBC) tunisien correspond un couple (racine, MBC) MSA et dans 333 cas, il lui en correspond plusieurs.

Nous reviendrons plus en détails sur l'ambiguïté dans la partie 6.

5. Les lettres défectueuses dans l'arabe sont **و** [w] et **ي** [y]

6 Evaluation

Le processus de traduction d'une forme verbale en tunisien en une forme verbale MSA se décompose en trois étapes : l'analyse morphologique à l'aide de l'outil MAGEAD adapté au tunisien, le transfert lexical réalisé au niveau des racines grâce à un lexique MSA-tunisien et la génération de la forme verbale MSA grace à l'outil MAGEAD pour le MSA. Rappelons que chacune de ces étapes est réversible et que l'on peut symétriquement traduire une forme verbale MSA en une forme verbale en tunisien.

De manière plus précise, à partir d'un verbe source, MAGEAD produit toutes ses analyses possibles, chacune d'elles est composée d'une racine-source, d'un MBC-source et de différents traits morphologiques. Le couple (racine-source, MBC-source) permet de faire un accès au dictionnaire pour extraire un ou plusieurs couples (racine-cible, MBC-cible). Les traits morphologiques sont quant à eux conservés tels quels. Le processus est décrit dans la figure 1.

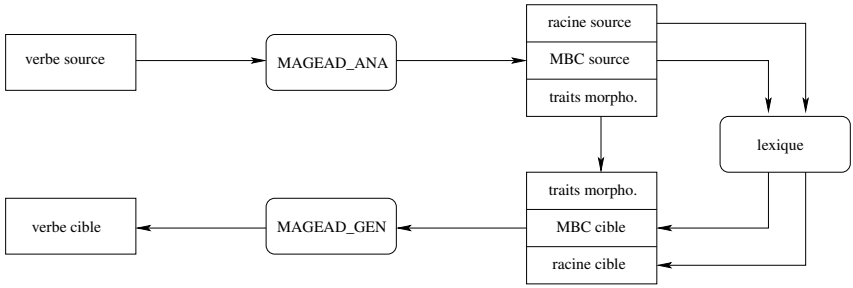

FIGURE 1: Traduction d'une forme verbale d'une langue source vers une langue cible

Cette architecture recèle deux sources d'ambiguïté. D'une part, l'analyse peut créer plusieurs couples (racine-source, MBC-source) et, d'autre part, le lexique peut proposer pour un couple (racine-source, MBC-source) plus d'un couple (racine-cible, MBC-cible).

Comme nous l'avons mentionné dans l'introduction, l'objectif général de ce travail n'est pas de produire un système de traduction du tunisien vers le MSA mais de générer à partir d'un texte tunisien une version de ce dernier sous une forme se rapprochant du MSA, de sorte que des outils de traitement automatique du MSA, tel que des étiqueteurs morpho-syntaxiques ou des analyseurs syntaxiques puissent être utilisés sur cette nouvelle forme du texte avec des résultats satisfaisants. La réelle évaluation sera donc réalisée sur la sortie de ces outils.

Les expériences décrites ici ne fournissent qu'une évaluation partielle, elles permettent de mesurer dans quelle mesure, pour une forme verbale tunisienne en entrée, la forme verbale MSA correcte est générée en sortie.

L'évaluation de ce processus est confronté au problème de l'absence de ressources écrites pour les dialectes. Afin de pallier ce problème, nous avons eu recours au livre (Dhouib, 2007) qui est une pièce de théatre écrite en tunisien. Les 1500 occurrences de formes verbales ont été identifiées et traduites en contexte, en tunisen, par deux locuteurs natifs. A l'issue de ce processus, 1500 couples (forme tunisienne, forme MSA) ont été produits et cet ensemble a été divisé en deux parties égales. La première constituant un ensemble de développement et la seconde un ensemble

de test. Deux métriques standard ont été utilisées pour évaluer le processus : la précision, qui indique la proportion de cas pour lesquel la forme cible correcte a été produite et l'ambiguïté, qui indique le nombre de formes cible produites en moyenne, pour une forme source.

Les expériences ont été réalisées dans le sens tunisien vers MSA et dans le sens MSA vers tunisien. Nous avons distingué les résultats sur les types et sur les occurrences. L'ensemble de développement a permis de combler quelques lacunes de l'analyseur et du générateur morphologique et d'enrichir notre lexique. Les résultats des expériences sur le corpus de développement sont donnés dans le tableau 5.

	précision		ambiguïté	
	occurences	types	occurences	types
TUN ⇒ MSA	87.65	86.68	25.42	23.33
MSA ⇒ TUN	89.56	88.74	1.25	2.87

TABLE 5: Précision et ambiguïté de la traduction des verbes de l'ensemble de développement

Ces expériences ont été, ensuite, lancées sur l'ensemble de test (cf. tableau 6). La grande différence entre l'ensemble de développement et celui du test est le lexique. En effet, dans les expériences sur les données de développement, toutes les paires (racine, MBC) qui ne se trouvent pas dans le lexique ont été rajoutées.

	précision		ambiguïté	
	occurences	types	occurences	types
TUN ⇒ MSA	76.43	74.52	26.82	25.57
MSA ⇒ TUN	79.24	75.1	1.47	3.1

TABLE 6: Précision et ambiguïté de la traduction des verbes de l'ensemble de test

Une analyse d'erreurs dans le sens TUN⇒MSA a montré que 34.6% des erreurs proviennent du lexique, alors que 14.5% d'erreurs proviennent de MAGEAD MSA et 51.9% proviennent de MAGEAD tunisien. La plupart des erreurs commises par MAGEAD sont dûes aux phénomènes morphologiques qui n'ont pas encore été implémentés, en particulier les verbes quadrilitères et l'impératif des verbes défectueux. D'autres erreurs spécifiques à MAGEAD tunisien proviennent des verbes pour lesquels la première ou la troisième lettre de la racine est "hamza" ء ['] qui nécéssitent un traitement spécifique. D'autre part, cette analyse d'erreurs a révélé deux types d'ambiguïtés : l'ambiguïté lexicale, dans 30% des cas et l'ambiguïté morphologique dans 70% des cas.

7 Conclusion

Nous avons proposé dans cet article un système de traduction de formes verbales depuis le tunisien vers le MSA et vice-versa. Ce travail s'inscrit dans un projet plus général de traduction des dialectes de l'arabe vers des approximations du MSA. Les résultats donnés par ce système sont environ 76% pour le passage du dialecte tunisien à l'arabe standard et 79% de performances dans l'autre sens.

L'architecture développée va être utilisée pour traduire les noms. Nous n'avons pas traité ici

le problème de l'ambiguïté : comment choisir une traduction lorsque plusieurs sont proposées par le système ? Il sera traité dans une étape ultérieure par l'utilisation d'un modèle de langage, appris sur des corpus MSA. Un tel modèle de langage permettra de sélectionner la séquence de meilleure probabilité.

Références

ABDILLAHI, N., NOCERA, P. et TORRES-MORENO, J. (2006). Boîtes à outils tal pour les langues peu informatisées : le cas du somali. *Actes de JADT*, 6:697–705.

ALTABBAA, M., AL-ZARAEE, A. et SHUKAIRY, M. (2010). An arabic morphological analyser and part-of-speech tagger. *Actes de JADT*, page 50.

BUCKWALTER, T. (2004). Buckwalter arabic morphological analyser version 2.0. *In Linguistic Data Consortium, University of Pennsylvania. LDC Cat alog No. :LDC2004L02, ISBN 1-58563-324-0*.

DHOUIB, E. (2007). El makki w zakiyya. Maison d'édition manshuwrat manara, Tunis.

FERGUSON, C. (1959). Diglossia. *Word*, 15(2).

HABASH, N. et RAMBOW, O. (2006). Magead : a morphological analyzer and generator for the arabic dialects. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

HABASH, N., RAMBOW, O. et KIRAZ, G. (2005). Morphological analysis and generation for arabic dialects. *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24. Association for Computational Linguistics.

KIRAZ, G. (2000). Multitiered nonlinear morphology using multitape finite automata : a case study on syriac and arabic. *Computational Linguistics*, 26(1):77–105.

MAAMOURI, M., BIES, A., BUCKWALTER, T. et MEKKI, W. (2004). The penn arabic treebank : Building a large-scale annotated arabic corpus. *In NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

MEJRI, S., MOSBAH, S. et SFAR, I. (2009). Pluringuisme et diglossie en tunisie. *Synergies Tunisie n 1*, pages 53–74.

OUERHANI, B. (2009). Interférence entre le dialectal et le littéral en tunisie : Le cas de la morphologie verbale. *Synergies Tunisie n 1*, pages 75–84.

SCHERRER, Y. *et al.* (2009). Un système de traduction automatique paramétré par des atlas dialectologiques. *Actes de TALN*.

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

SHAALAN, K., BAKR, H. et ZIEDAN, I. (2007). Transferring egyptian colloquial dialect into modern standard arabic. *In International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria*, pages 525–529.

SMRŽ, O. (2007). Elixirfm : implementation of functional arabic morphology. *In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages : Common Issues and Resources*, pages 1–8. Association for Computational Linguistics.