Génération de corpus en dialecte tunisien pour l'adaptation de modèles de langage

Rahma Boujelbane[1,2]
(1) ANLP_MIRACL, Sfax, Tunisie
(2) LIF, UMR7279, 13288, Marseille, France
Rahma.boujelbane@gmail.com

RÉSUMÉ

Ces derniers temps, vu la situation préoccupante du monde arabe, les dialectes arabes et notamment le dialecte tunisien est devenu de plus en plus utilisé dans les interviews, les journaux télévisés et les émissions de débats. Cependant, cette situation présente des conséquences négatives importantes pour le Traitement Automatique du Langage Naturel (TALN): depuis que les dialectes parlés ne sont pas officiellement écrits et n'ont pas d'orthographe standard, il est très coûteux d'obtenir des corpus adéquats à utiliser pour des outils de TALN. Par conséquent, il n'existe pas des corpus parallèles entre l'Arabe Standard Moderne(ASM) et le Dialecte Tunisien (DT). Dans ce travail, nous proposons une méthode pour la création d'un lexique bilingue ASM–DT et un processus pour la génération automatique de corpus dialectaux. Ces ressources vont servir à la construction d'un modèle de langage pour les journaux télévisés tunisiens, afin de l'intégrer dans un Système de Reconnaissance Automatique de Parole (SRAP).

ABSTRACT

Generation of tunisian dialect corpora for adapting language models.

Lately, given the serious situation in the Arab world, the Arab dialects such as Tunisian dialect became increasingly used and represented in the interviews, news and debate programs. However, this situation presents negative consequences for Natural Language Processing (NLP): Since dialects are not officially written and have no orthographic standard, it is very costly to obtain adequate corpora to train NLP tools. Therefore, it does not even exist parallel corpora between Standard Arabic (MSA) and Tunisian Dialect (TD). In this work, we propose a method for the creation of a bilingual lexicon MSA-TD and an automatic process for generating dialectal corpora. These resources will be used to build a language model for Tunisian news, in order to integrate it into an Automatic Speech Recognition (ASR).

MOTS-CLÉS : Dialecte Tunisien, lexique ASM-DT, TDT: Tunisian Dialect Translator.
KEYWORDS : Tunisian Dialect, MSA-TD lexicon, TDT: Tunisian Dialect Translator.

1 Introduction

L'utilisation de corpus constitue un problème crucial pour les langues disposant peu de ressources électroniques et peu informatisées comme de nombreux dialectes arabes. En effet, la construction de corpus est une étape capitale pour une bonne réalisation d'outils de traitement automatique de la langue tels que les systèmes de reconnaissance de parole qui nécessitent des données textuelles en grande quantité pour apprendre le vocabulaire d'une langue. Récemment en Tunisie, la révolution a touché non seulement le peuple

mais aussi les médias. Par conséquent, en un an tout le paysage médiatique a été bouleversé: les chaines, les émissions de débats et les journaux télévisés se sont multipliés. Ceci a donné naissance à un nouveau type de discours médiatique. En effet, la majorité des discours ne sont plus en ASM mais ils présentent une alternance entre le ASM et le dialecte. En effet, nous pourrons distinguer dans un même discours des mots en ASM, des mots en DT et des mots ASM «dialectalisés» tel qu'un mot avec une racine ASM et des affixes dialectales. Face à cette situation un SRAP conçu pour le ASM serait incapable de transcrire cette nouvelle langue. Pour cela, nous focalisons dans le présent travail à construire des ressources représentatives de ce mélange entre le ASM et le DT. Pour ce faire, nous proposons une approche basée sur deux étapes principales: La première consiste à construire une base lexicale, dans laquelle nous introduisons des règles de correspondance entre des structures en arabe standard et des structures dialectales. La deuxième étape se repose sur l'exploitation de cette base lexicale afin de générer des corpus dialectaux.

2 Travaux connexes

Les langues orales qui n'ont pas de forme écrite répandue peuvent être classées comme des langues peu dotées. De ce fait, plusieurs travaux ont tenté de pallier les problèmes liés à l'informatisation des langues peu dotées. **(Y. Scherrer, 2008),** dont le but d'informatiser le dialecte existant en Suisse, a développé un système de traduction allemand standard-suisse allemand. Le système développé traduit en se basant, sur un lexique bilingue, l'allemand standard vers n'importe quelle variété du continuum dialectal de la Suisse alémanique. Par ailleurs, les auteurs dans **(Shaalan et al., 2007**) se sont intéressés à l'informatisation du dialecte égyptien, l'une des variantes de l'arabe standard. Les auteurs ont proposé un système de traduction dialecte EGYptien EGY-ASM. A cette fin, ils ont essayé de construire un corpus parallèle EGY-ASM, ceci en se basant sur des règles de correspondance EGY- ASM. A part les dialectes, il existe plusieurs langues parmi la famille des langues peu dotées qui n'ont pas de relation avec une langue bien dotée comme le somalien et le khmer, etc. Ainsi, l'approche proposée pour constituer des corpus en somalien dans **(Nimaan et al., 2006)** se repose sur plusieurs scénarios: collecte de corpus à partir du Web, synthèse automatique de textes et traduction automatique français-somali. Et pour solliciter le manque de ressources en khmer **(Seng, 2010)** a choisi les sites de nouvelles en khmer au fort contenu rédactionnel pour collecter les corpus textuels. La revue de la littérature nous a montré qu'il n'y a pas assez de travaux qui ont traité l'arabe tunisien, la langue cible de ce travail. Le travail de **(Graja et al, 2011**) par exemple traite le dialecte tunisien pour la compréhension de parole. Cependant, ce travail utilise un domaine limité à savoir le transport ferroviaire où le vocabulaire est assez limité. Pour disposer des données, les chercheurs ne se sont basés que sur des transcriptions manuelles de conversations entre l'agent du guichet et les voyageurs. Or, un vocabulaire limité pose un problème si nous souhaitons modéliser un modèle de langage pour un système de reconnaissance des émissions de la télévision ayant un vocabulaire large et varié.

3 Ressources pour le dialecte tunisien

Le dialecte tunisien est une langue arabe rattachée à l'arabe maghrébin parlée par douze

millions de personnes vivant principalement en Tunisie. Bien que la langue officielle soit l'arabe littéral, il est généralement connu de ses locuteurs sous le nom de 'Darija' ou 'Tounsi' ce qui signifie tout simplement «tunisien», afin de le distinguer de l'arabe littéral (Baccouche, 1994). Dans les deux dernières années, ce dialecte est devenu la langue parlée dans la plupart des médias au lieu de l'arabe standard. Mais, cette forme dialectale a une forme sophistiquée: elle présente des formes mixtes ASM-DT et elle est en même temps un dialecte et une langue proche du ASM. Ainsi, étant donné l'écart faible entre cette forme dialectale et le ASM, les ressources disponibles pour le ASM peuvent être avantageusement utilisées pour créer des ressources dialectales.

3.1 L'Arabic TreeBank pour la création d'un lexique bilingue ASM-DT

Au début de cette étude et lors d'une convention avec le LDC (Linguistic Data Consortium), nous avons eu l'opportunité de travailler sur le corpus Arabic TreeBank ATB (Maamouri, 2004). Il s'agit d'un corpus contenant 120 transcriptions d'émissions d'actualité en arabe standard diffusées par différentes chaines arabes. Le corpus transcrit contient 51 080 mots annotés morpho-syntaxiquement et syntaxiquement. Pour créer un lexique en dialecte tunisien, nous avons essayé de construire en partant de l'ATB un lexique de traduction ASM-DT. Pour ce faire, nous avons adopté une méthode de transformation, du ASM vers le dialecte, basée sur les parties du discours des mots de l'ATB. Ceci permettra d'obtenir non seulement des dictionnaires bilingues mais aussi un ATB en tunisien utile pour des applications TALN (Figure 1). Nous expliquons dans ce travail les modèles de transformation ainsi que les structures des dictionnaires que nous avons définis pour les verbes et les différents outils syntaxiques de l'ATB.

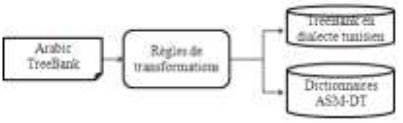

FIGURE 1 – Méthode pour la création de ressources en dialecte tunisien

3.2 Construction d'un lexique pour les verbes

Comme nous visons à adapter les outils ASM au dialecte tunisien, nous avons essayé de construire pour les verbes en DT les mêmes concepts que ceux d'ASM. En arabe, les principaux concepts verbaux sont:

1-Lemme: Il s'agit d'un concept fondamental dans l'analyse des textes. Les mots arabes peuvent être analysés comme étant une racine insérée dans un modèle constituant ainsi les lemmes de mots. Les verbes dans l'ATB sont présentés sous leur formes fléchies, nous avons extrait leurs lemmes et leurs racines en utilisant l'analyseur morphologique ELEXIR FM développé par (Smrž, 2007). Étant donné que nous sommes des locuteurs natifs du dialecte tunisien, nous avons construit manuellement à chacun des lemmes ASM des lemmes en dialecte. En résultat, nous avons constaté que 60% des verbes se comportent différemment en passant vers le dialecte ce qui prouve la différence entre le couple ASM-DT. Ainsi, ayant 1500 lemmes en DT et partant du fait que les verbes en ASM possèdent des schèmes décrivant leurs comportements morphologiques lors de la

conjugaison, nous avons cherché à attribuer des schèmes aux verbes en DT.

2-Pattern: Les patterns ou les schèmes en ASM sont des modèles avec différentes structures qui sont appliquées à la racine pour créer un lemme. Pour chaque racine, nous pouvons appliquer différents modèles pour avoir des lemmes avec des significations différentes. Le challenge dans la construction des schèmes pour les verbes dialectaux consiste à trouver des modèles similaires à ceux en ASM. Ainsi, en étudiant la morphologie des lemmes dialectaux, nous avons remarqué qu'il est possible d'attribuer aux lemmes en DT les mêmes modèles que ceux en ASM mais en définissant en plus d'autres modèles qui seront des schèmes fils pour les schèmes de base. En fait, ce processus a permis de distinguer 32 schèmes pour les verbes en DT, alors qu'il y avait 15 en ASM. Cela est dû à la richesse morphologique des lemmes dialectaux. Par exemple:

En ASM: le verbe $Arak/ شارك(forme au passé) yu$Arik/participerيشارك(forme au présent) ; et دافع/dAfaE (forme au passé) يدافع/yudAFiE/défendre(forme au présent) appartiennent au ASM-pattern-II: CvACvC (forme au passé)/yvCvACvC (forme au présent). Notons que les voyelles sont les mêmes pour les deux verbes. En passant vers le dialecte, les racines ainsi que les modèles de ces verbes restent les mêmes CvACvC/yvCACvC mais la voyelle de la seconde consonne n'est plus la même pour les deux verbes. Or en ASM, la marque de cette voyelle est un critère fondamental pour classer un verbe sous un pattern (Ouerhani, 2009) c'est pourquoi nous avons proposé de définir des sous-patterns pour le pattern II, et ce en divisant le pattern-II en pattern-II-i: CACiC/yvCACiC et pattern-II-a:Cacac/yvCACaC. Par conséquent, $Arak/yu$Arik qui devient en DT $Arik/yi$Arik/ va appartenir au pattern-II-i: CACIC/yiCACiC et dAfaE/yudAFiE qui devient en DT dAfaE/yidAFaE va appartenir au pattern-II-a: CACAC / yiCACaC. Donc, en adoptant ce raisonnement, nous avons réussi avec les verbes de l'ATB à définir des schèmes pour les verbes en DT.

3-Racine: Elle est la source fondamentale de toutes les formes des verbes arabes. La racine n'est pas un vrai mot, il s'agit plutôt d'une séquence de trois consonnes qui peut être trouvée dans tous les mots qui lui sont liés. La plupart des racines sont composées de trois lettres, très peu sont de quatre ou cinq consonnes. En dialecte tunisien, il n'existe pas encore de standard pour la définition de la racine. Pour cela, la construction de racine en dialecte n'est pas évidente, surtout quand le verbe change complètement de racine en passant du ASM vers le dialecte. En fait, pour définir une racine pour les verbes TUN, nous avons adopté une méthode déductive. En effet, la règle en ASM dit que racine + schème = lemme (1). Dans notre cas, nous avons déjà défini le lemme TUN et le schème TUN. En suivant la règle (1), l'extraction de la racine est rendue alors facile. Par exemple, nous avons classé le lemme إِسْتَنَّى /Aistan~aY/Attendre dans le schème AistaCCaC

Racine (?) + AistaCCaC = إِسْتَنَّى /Aistan~Y

En suivant (1), la racine est alors «ننى » [nnY]. En fait, nous pouvons dire que la définition des racines est une question problématique et qui pourrait admettre plus de discussion. D'après la démarche adoptée, c'est comme si, nous avons forcé la racine à être [nnY]. En effet, si nous classons إِسْتَنَّى /Aistann~aY sous le schème AiCtaCaC, la racine dans ce cas doit être سنن /snn. La racine peut être aussi quadrilatère سنني /snnY si nous classons إِسْتَنَّى /Aistann~aY sous le schème AiCCaCaC. Mais comme il n y'a pas de

standard, nous avons fait de notre mieux pour être le plus logique possible en définissant la racine dialectale.

3.3 Modélisation des concepts verbaux dialectaux dans le lexique ASM-DT

Les différentes transformations verbales décrites ci-dessus, sont modélisées et stockées dans un dictionnaire de verbes de la manière suivante: chaque bloc verbal ASM, contenant le lemme-ASM, schème-ASM et la racine-ASM, lui correspond respectivement un bloc DT contenant le lemme-TUN, la racine-TUN et le schème-TUN. La connaissance du bloc TUN nous permet de définir automatiquement les différentes formes fléchies du verbe TUN. La Figure 2 décrit la structure que nous avons définie pour stocker une unité verbale ASM-DT.

```
<DIC_TUN_VERBS_FORM>
 <LEXICAL-ENTRY POS="VERB">
<VERB ID-VERB="48">
    <ASM-LEMMA>
     <Headword-ASM>عَايَنَ</Headword-ASM>
      <Pattern>فاعل</Pattern>
      <Root-ASM>عين</Root-ASM>
      <Gloss lang= "fr" > Observer</Gloss>
    </ASM-LEMMA>
<TUN-VERB Sense= "1" >
   <Cat-Tun-Verb Category= "TUN--VERB--I--au--yi" />
<Root-Tun-Verb>شوف</Root-Tun-Verb>
<Conjug-Tun-Verb>
<TENSE>
<FORM Type= "IV" >
<VOICE Label="Active">
<Features Val_Number_Gender="1S">
<Verb_Conj>نشُوفْ</Verb_Conj>
<Struct-Deriv>Ø+شوف+ن</Struct-Deriv>
</Features>
</VOICE>
<VOICE Label="Passive">
      :::
</VOICE>
</Form >
<FORM Type= "PV" >
  <VOICE Label="Active">
       ::
</VOICE>
 <VOICE Label="Passive">
::::
</VOICE>
</Form >
<FORM Type= "CV" >
 <FeaturesVal_ Number _Gender="2S">
<Verb_Conj>شُوفْ</Verb_Conj>
        <Struct-Deriv>Ø+شوف+Ø</Struct-Deriv>
       </Features>
   </FORM>
 </TENSE>
 </Conjug-Tun-Verb
</TUN-VERB>
 </LEXICAL-ENTRY>
</DIC_TUN_VERBS_FORM>
```

FIGURE 2 – Structure de stockage d'une unité verbale

3.4 Règles de transformation pour la traduction des mots outils:

3.4.1 Transformation dépendante du contexte

Pour ce type de transformation, nous avons proposé de décrire les différents contextes dont peut dépendre un mot outil sous forme de règles. Nous désignons par une règle basée-contexte, le passage ASM-DT qui s'appuie sur des règles de transformation. En effet étant donnée un mot MK, on dit que la transformation de MK se base sur le contexte s'il donne une nouvelle traduction à chaque fois qu'on lui change le contexte. RT_k: $X + M + Y = TD_k$

X = $\sum_{j=1}^{m}$ Mj: POSj ; Y = $\sum_{i=1}^{n}$ Mi: POSi ; k varie de 1 à z ;

RT_k: Règle de transformation n°$_k$; POS : Partie de discours ; M: Mot outil ; TDk: Traduction n°$_k$.

Pour chaque mot outil, plusieurs configurations peuvent se présenter donnant à chaque fois une nouvelle traduction. La transformation d'un mot outil peut dépendre soit des mots qui le précèdent (X), soit qui le suivent (Y), soit des deux. Si aucun contexte ne se présente alors une traduction par défaut sera affectée au mot outil. Prenons l'exemple de la particule « حَتَّى »/HatY /pour que/ qui possède la POS: sub-conj dans l'ATB. Pour cette particule nous avons développé des règles conformément à trois contextes différents vus dans l'ATB.

1- HatY/حَتَّى + verb = باش/bA$ (TUN-particle) + TUN_verb (DIC-TUN-Verb)

2–HatY/حَتَّى + NEG_PART = bA$/باش(TUN-particle) + TUN_NEG_PART(DIC-TUN-NEG_PART).

Sinon

3- HatY/حَتَّى = HatY/حَتَّى (dans tous les autres contextes)

Le tableau 1 montre la manière dont on représente une règle dans le lexique. En effet, pour chaque mot outil nous avons défini un ensemble de contextes, chaque contexte contient une ou plusieurs configurations. La configuration décrit la position et la partie de discours du mot par rapport au mot outil. Chaque contexte lui correspond une traduction en tunisien. La traduction peut être soit directe soit indirecte c.à.d. elle fait appel à un autre dictionnaire de notre base lexicale (autre que celui du mot outil concerné).

Règle de transformation	HatY/حَتَّى + verbe = باش (TUN-particle) + TUN_verbe
Représentation de la règle dans le dictionnaire	<SUB_Conj ID="10"> <ASM-LEMMA>حَتَّى</ASM-LEMMA> <GLOSS lang="fr">Jusqu'à ce que / à</GLOSS> <CONTEXT ID="1"> <CONFIG ID= "1" Position="Après" POS="Verb" /> <TOKEN> <TUN ID="1">باش </TUN> <TUN ID="2"> </TUN> <TUN ID="3" DIC= « verbs » POS="verb" /> </TOKEN> </CONTEXT> …… <CONTEXT ID="3"> ….. </Sub_Conj>

TABLE 1 – Structure de stockage d'une transformation dépendante du contexte

3.4.2 Transformation syntaxique

Les transformations qui requièrent le changement de l'ordre syntaxique est une catégorie de transformation qui est aussi dépendante de contexte. Il s'agit de changer l'ordre des mots pour qu'ils aient un sens en dialecte. Dans notre travail, nous avons traité le niveau syntaxique au niveau de quelques groupes nominaux tels que:

ASM: كُتُب كَثِيرة /kutubun kavirap/ Noun + ADJ

DT: برشا كتُب/ bar$A ktub/ ADJ + Noun

L'intérêt ici, consiste à montrer la faisabilité de ce type de transformation et qu'on pourra intégrer dans notre base lexicale d'autres règles de ce type. Nous pourrons penser par exemple à changer les structures VSO (Verbe Sujet Object) en des structures SVO (Sujet Verbe Objet) puisqu'elles sont fréquemment utilisés en dialecte (Baccouche, 2003). Le tableau 2 illustre le stockage d'une règle contentant une transformation d'ordre syntaxique.

Transformation syntaxique	ASM: Noun+ ADJ -> DT: ADJ+Noun
Représentation de la règle dans le dictionnaire	<Noun-ASM ID="5"> <ASM-LEMMA> كُتُب </ASM-LEMMA> <GLOSS lang="fr">livres</GLOSS> <CONTEXT ID="1"> <CONFIGID="1"Position="Après"POS="ADJ" /> <TOKEN> <TUNID="1"DIC="ADJECTIVES" POS="ADJ" /> <TUN ID="2" /> <TUN ID="3">كتُب </TUN> </TOKEN> </CONTEXT>

TABLE 2 – Structure de stockage d'une transformation syntaxique

L'étude des différents contextes de mot outils nous a permis de développer 316 règles. Le tableau 3 montre le nombre de règles développés pour chaque mot outils.

	Préposition	Conjonction	Pseudo-verbe	Adverbe	Pronom	Particule	Interjection
Occurrence_ ASM	49924	36498	1505	1662	1642	6245	38
Mot différent_ASM	13	23	7	36	44	23	6
Nbre de règles	141	42	24	45	53	51	6

TABLE 3 – Statistique des règles de transformation développées

4 Génération automatique des corpus en dialecte tunisien

A fin d'assurer le recueil du maximum possible de ressources, nous avons développé un outil baptisé Tunisian Dialect Translator (TDT). Ce dernier est capable de générer automatiquement des textes en tunisien en exploitant le lexique bilingue développé et de l'enrichir. Le TDT fonctionne selon la démarche suivante :

1-Etiqueter morphosyntaxiquement un texte ASM: Chaque corpus textuel en ASM est analysé morphosyntaxiquement à l'aide de l'analyseur MADA (Morphological Analyser and disambiguisator of Arabic Dialect) (Habash, 2010). Il s'agit d'un outil multitâche: Il effectue à la fois la segmentation, la discrétisation, la lemmatisation, l'analyse morphologique et l'étiquetage morphosyntaxique. L'apport principal de cet outil est la désambigüisation.

2- Exploiter la base-lexicale ASM-DT: En se basant sur chaque partie du discours résultant de l'étiquetage de MADA, nous exploitons la base lexicale ASM-dialecte que nous avons développée et ce en créant pour chaque structure ASM sa traduction correspondante en DT.

3-Enrichir le lexique: le texte obtenu de l'étape précédente n'est pas toujours traduit parfaitement, vu que la base lexicale ne couvre pas tous les mots. Pour cela, dans le but d'améliorer la qualité de la traduction et d'enrichir davantage notre lexique, nous avons développé un module d'enrichissement semi automatique. Il permet de filtrer tous les mots ASM pour lesquels une traduction n'a pas été fournie. Ces mots sont intégrés d'une manière semi-automatique dans le lexique après avoir proposé les traductions correspondantes. La figure 3 illustre la démarche décrite.

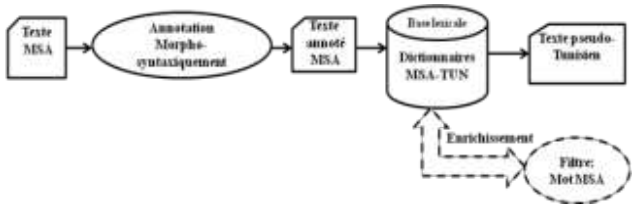

FIGURE 3 – Méthode pour la génération automatique de corpus en dialecte tunisien

5 Évaluation

Dans cette section, nous présentons une évaluation pour le lexique développé. En effet, nous avons demandé à des juges, qui sont des locuteurs natifs du dialecte tunisien, de proposer des traductions à des unités lexicales prises de notre dictionnaire.

5.1 Évaluation du lexique des verbes

Pour évaluer les verbes, nous avons proposé dans un premier temps à des juges de nous traduire un échantillon de verbes. En effet, parmi les 1500 verbes, nous avons pris aléatoirement un échantillon contenant 150 verbes. L'échantillon comporte 52 verbes qui ne changent pas du ASM vers le dialecte et 98 verbes qui changent complètement. Nous avons demandé par la suite à 47 juges de proposer des traductions à ces verbes. Les pourcentages calculés traduisent le pourcentage d'accord pour chaque verbe entre les traductions des juges et la traduction proposée dans notre lexique. Le tableau 4 représente les résultats obtenus.

Verbes	Inchangés	Changés	Total
Nombre de verbes	52	98	150
Accord	97,17%	63,21%	74,97%

TABLE 4 – Évaluation des verbes

La baisse de l'accord au niveau des verbes inchangés est due au fait que nous n'avons pas pris en compte l'aspect sémantique en faisant la traduction des verbes. En effet, un verbe peut avoir plusieurs sens selon la phrase où il se trouve.

5.2 Évaluation du lexique des mots outils

Puisque la traduction de la majorité des mots outils dépend du contexte, nous avons donné à 5 juges 89 phrases contenant 133 mots outils. Les mots outils se répètent parfois dans les phrases mais diffèrent du contexte. Nous avons demandé aux juges de traduire seulement les mots outils.

	2 juges	3 juges	4 juges	5 juges
Accord	72,69%	74,53%	71,34%	71,23%
Désaccord Total	18,79%	15,03%	14,28%.	12,03%

TABLE 5 – Évaluation des mots outils

Le tableau (5) donne les pourcentages d'accord entre les traductions des juges et celles de notre outil. La variation des pourcentages est due au fait que pour quelques mots outils, les juges ne s'accordent pas entre eux-mêmes. Le tableau présente aussi les

pourcentages de désaccord total entre les juges et le système. Le désaccord total se présente lorsqu'aucun juge ne donne une traduction similaire à celle donnée par le système. En augmentant le nombre de juges, le désaccord diminue ce qui prouve que notre base lexicale est capable de générer des traductions acceptables par plusieurs juges.

6 Conclusion

Dans cet article, nous avons décrit un processus de création de lexique et de génération de texte en dialecte tunisien dans le but de créer des corpus textuels pour entrainer un modèle de langage d'un système de reconnaissance. Ce processus s'est déroulé en deux phases. Dans la première, nous nous sommes focalisés sur la création d'une base bilingue ASM-DT en partant de l'Arabic TreeBank. Cette base a été exploitée aussi dans un travail d'adaptation de MAGEAD [Habash et al 2006] (Morphological Analyser and Generator of Arabic Dialect) au dialecte tunisien. Ceci est bien expliqué dans (Hamdi et al., 2013). Dans la deuxième phase, nous avons exploité cette base pour automatiser la tâche de la génération des textes en dialecte tunisien. L'orientation future de ce travail consiste à enrichir la base lexicale afin d'élargir la couverture lexicale dialectale. Nous envisageons aussi à proposer un modèle de langage pour les corpus dialectaux qu'on a obtenus et les évaluer par rapport à des transcriptions réelles. Des expériences en cours de réalisation sur le modèle de langage pour ces types de corpus ont montré que l'intégration de ces nouveaux corpus peut influer avantageusement sur le modèle de langage.

Remerciements

Je tiens à témoigner ma sincère reconnaissance à l'étudiante de master Mlle.Siwar benAyed qui m'a aidée à réaliser ce travail. Je remercie également M .Fréderic Bechet, Mme Mariem Ellouze, et Mme Lamia Belguith pour leurs remarques précieuses et leur participation au cheminement de ce travail au sein du laboratoire ANLP MIRACL et LIF Marseille .

Références

BACCOUCHE, T. (1994). L'emprunt En Arabe Moderne, Beit Elhikma Et Iblv, Tunis.

BACCOUCHE, T. (2003). La langue arabe: spécificités et évolution.

DIKI-KIDIRI, M. (2007). Comment assurer la présence d'une langue dans le cyberespace, UNESCO, Paris.

GRAJA, M., JAOUA, M. ET HADRICH BELGUITH, L. (2011). Building ontologies to understand spoken Tunisian dialect, *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, pp.23-32.

HABASH, N., RAMBOW, O., ROTH, R. (2009). MADA + TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR),* Cairo, Egypt.

HAMDI, A., BOUJELBANE, R., HABASH, N., NASR, A., Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde, *Actes de TALN2013 (Traitement automatique des langues naturelles)*, Nante, France.

NIMAAN, A., NOCERA, P. ET TPRRES-MORENO, JM. (2006). Boîte à outils TAL pour des langues peu informatisées : le cas du somali. *JADT* 2006 (*Journées internationales d'Analyse statistique des Données Textuelles)*. France, pp.694-701.

MAAMOURI,M. ET BIES, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools, *Workshop on Computational Approaches to Arabic Script-based Languages*, COLING, Genève, Suisse.

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

SCHERRER,Y. (2008), Transducteurs à fenêtre glissante pour l'induction lexicale, *RECITAL*, Avignon, France.

SHAALAN, K., ABOUBAKR, HM., ET ZIEDAN,I. (2007). Transferring Egyptian Colloquial Dialect into Modern Standard Arabic. *RANLP (International Conference on Recent Advances in Natural Language Processing),* pp.525-529.Brovets, Bulgarie.

SMRŽ, O. (2007). Computational Approaches to Semitic Languages, ACL, Prague.

OUERHANI,B., (2009), Interférence entre le dialectal et le littéral en Tunisie: Le cas de la morphologie verbale, Synergies Tunisie n° 1, pp. 75-84, Tunisie.