Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora

Rahma Boujelbane
ANLP Research Group, MIRACL Lab, University of Sfax, Tunisia
rahma.boujelbane@gmail.com

Meriem Ellouze Khemekhem
ANLP Research Group, MIRACL Lab, University of Sfax, Tunisia
mariem.ellouze@planet.com

Lamia Hadrich Belguith
ANLP Research Group, MIRACL Lab, University of Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Abstract

Nowadays in tunisia, the arabic Tunisian Dialect (TD) has become progressively used in interviews, news and debate programs instead of Modern Standard Arabic (MSA). Thus, this gave birth to a new kind of language. Indeed, the majority of speech is no longer made in MSA but alternates between MSA and TD. This situation has important negative consequences on Automatic Speech Recognition (ASR): since the spoken dialects are not officially written and do not have a standard orthography, it is very costly to obtain adequate annotated corpora to use for training language models and building vocabulary. There are neither parallel corpora involving Tunisian dialect and MSA nor dictionaries. In this paper, we describe a method for building a bilingual dictionary using explicit knowledge about the relation between TD and MSA. We also present an automatic process for creating Tunisian Dialect (TD) corpora.

1 Introduction

Recently, due to the changes that have occurred in the Arab world, we noticed a new remarkable diversity in the media. The Arabic dialects used in daily life have become progressively used and represented in interviews, news and debate programs instead of Modern Standard Arabic (MSA). In Tunisia, for example, the revolution has affected not only the people but also the media.

For that reason, the media programs have been changed: television channels, political debates and broadcasts news have been multiplied. This gave birth to a new kind of language. Indeed, the majority of speech is no longer made in MSA but alternates between MSA and Tunisian Dialect (TD). Thus, we can distinguish in the same speech, MSA words, TD words and MSA-TD words such as a word with an MSA component (root) and dialectal affixes. This situation poses significant challenges to NLP: In fact, applying NLP tools designed for MSA to TD yields a significantly lower performance, making it imperative to direct research towards building resources and tools that make it possible to process this kind of language. In our case, we aim to convert this new language into text. However, this process presents a series of linguistic and computational challenges. Some of these relate to language modeling: studying large amounts of text to learn about patterns of words in a language. This task is complicated because of the total lack of TD resources, whether parallel TD-MSA text or dictionaries. In this paper, we describe a method that helps to create Tunisian Dialect (TD) text corpora and the associated lexical resources and also build a bilingual MSA-TD dictionary. This paper is organized as follows: After discussing related work, we present our method to deal with the lack of Tunisian resources (Section3). We then proceed to discuss the method in details: we explain the manner of creating Tunisian verbal

International Joint Conference on Natural Language Processing, pages 419–428,
Nagoya, Japan, 14-18 October 2013.

resources (Sections 4 and 5). We present in Section 6 a tool for generating dialectal corpora. We evaluate and discuss the results in Section 7.

2 Related work

Arabic dialects have earned the status of living languages in linguistic studies, thus we see the emergence of a serious effort to study patterns and regularities in these linguistic varieties of Arabic (Brustad, 2000; Holes, 2004; Erwin, 1963).

To date, most of these studies have been field studies or theoretical in nature with limited annotated data. In fact, Dialectal Arabic (DA) is emerging as the language of the news and of many varieties of television programs, and also of informal communication online, in emails, blogs, discussion forums, chats, SMS, etc. In current statistical Natural Language Processing (NLP) there is an inherent need for large scale annotated resources for a language (Diab *et al.*, 2010).

But, research on computerization of DA is still in its early stages especially for TD. Several researchers have explored the idea of exploiting existing MSA rich resources to build tools for DA NLP. For example, (Chiang *et al.*, 2006) built syntactic parsers for DA trained on MSA Treebanks. Such approaches typically expect the presence of tools/resources to relate DA words to their MSA variants or translations. Given that DA and MSA do not have much parallel in terms of corpora to help translate DA-to-MSA, (Abo Bakr *et al.*, 2008) introduced a hybrid approach to transfer a sentence from Egyptian Arabic into MSA. This hybrid system consisted of a statistical system for tokenizing and tagging, and a rule-based system for constructing diacritized MSA sentences. Moreover, (Al-Sabbagh and Girju, 2010) described an approach of mining the Web to build an Egyptian-to-MSA lexicon. (Diab *et al.*, 2010) presented an information retrieval project COLABA (Cross Lingual Arabic Blog Alerts) that aims to create resources and processing tools for dialectal Arabic blogs. The COLABA system consists in taking an MSA query and translating it or its component words into DA or alternatively converting all DA documents in the search collection into MSA before searching on them with the MSA query. To do so, they created DIRA (Dialectal Information Retrieval for Arabic), which is a term expansion tool for information retrieval over dialectal Arabic collections, especially the Egyptian and the Levantine dialects, using Modern Standard Arabic queries. (Habash and Rombow, 2006) presented MAGEAD (Morphological Analyser and Generator of Arabic dialect). MAGEAD works both on analyzing and generating Egyptian and Levantine verbs. The limitation of MAGEAD is that it doesn't deal with verbs that change their roots when moving from MSA to Dialect.

(Shaalan *et al.* 2007) proposed a system for translating MSA into the Egyptian dialect. To do so, they tried to build a parallel corpus between the Egyptian dialect and MSA based on mapping rules EGY-MSA.

As a conclusion, for MSA and its dialects, there are no naturally occurring parallel corpora. It is this fact that has led researchers to investigate the use of explicit linguistic knowledge.

Dialects are under-resourced languages:

Spoken languages which have no written form can be classified as under-resourced languages and as a consequence have no annotated resources. Therefore, several studies have attempted to overcome the problems of lack of resources for these languages. In order to computerize the existing Swiss dialect, (Scherrer, 2008) developed a translation system: standard German to Swiss German. The system developed is based on translating a bilingual lexicon from standard German to any variety of the dialect continuum of German-speaking Switzerland. Moreover, there are several languages from the group of under-resourced languages that do not have a relation with a well-resourced language. Indeed, (Nimaan *et al.* 2006) presented several scenarios to collect corpora in order to automatically process the Somali language: collecting a corpus from the Web, automatic synthesis of texts and machine translation of French into Somali. (SENG, 2010) selected news sites in Khmer to collect data in order to solicit the lack of resources in Khmer.

Related work vs. the Tunisian dialect: The literature shows that there is little work that dealt with the Tunisian dialect, the target language of this work. (Graja *et al.*, 2011) for example, treated the Tunisian dialect for understanding speech. To do so, the researchers relied on manual transcripts of conversations between agents at the train station and travelers. The scope of application is limited and so, the vocabulary is not very rich. However, a limited vocabulary is a problem if we want to model a language model for a system of recognition of

television programs with a wide and varied vocabulary. In addition, (Zribi *et al.*, 2013) presented OTTA (Orthographic Transcription for Tunisian Arabic), a set of guidelines orthography to transcribe Tunisian Arabic. This work is helpful for our case in that it will facilitate the identification of the orthography of the Tunisian words that we will build.

3 Method to create a Tunisian Dialect lexicon

In Arabic, there are almost no parallel corpora involving the Tunisian Dialect and MSA. Therefore, Machine Translation (MT) is not easy, especially when there are no MT resources available such as a naturally occurring parallel text or a transfer lexicon. So, to deal with this problem, we propose to leverage the large amount of annotated MSA resources available by exploiting MSA/dialect similarities and addressing known differences. Our approach consists first in studying the morphological, syntactic and lexical differences by exploiting the Penn Arabic Treebank. Second, we present these differences by developing rules and building dialectal concepts. Finally, we define a lexical data base to store these transformations into dictionaries**.**

3.1 Tunisian Dialect Vs. MSA

The Tunisian Arabic dialect is attached to the Arab Maghreb and is spoken by twelve million people living mainly in Tunisia. It is generally known to its speakers as the 'Darija' or 'Tounsi' which simply means "Tunisian", to distinguish it from Modern Standard Arabic (Baccouche, 1994).

The Tunisian dialect is considered as an under-resourced language. It has neither a standard orthographic or written text nor dictionaries.

Actually, there is no strict separation between Modern Standard Arabic (MSA) and its dialects, but a continuum dominated by mixed forms (MSA-Dialect). In the last two years, this dialect became the language spoken in most of the media instead of standard Arabic. But this dialect has a sophisticated form which mixes MSA and TD forms. Thus, given the similarities between TD and MSA, the resources available to MSA can be advantageously used to create dialectal resources.

3.2 Penn Arabic Treebank corpora to create a bilingual MSA-TD lexicon

Treebanks are important resources that allow for important research in general NLP applications. In the case of Arabic, two important treebanking efforts exist: the Penn Arabic Treebank (PATB) (Maamouri *et al.*, 2004; Maamouri *et al.*, 2009) and the Prague Arabic Dependency Treebank (PADT) (Smrž *et al.*, 2008). The PATB provides tokenization, complex POS tags, and syntactic structure; it also provides empty categories, diacritizations, and lemma choices. The ATB consists of 23,611 parse-annotated sentences (Bies and Maamouri, 2003; Maamouri and Bies, 2004) collected from Arabic newswire texts in Modern Standard Arabic (MSA). The ATB annotation scheme involves 497 different POS-tags with morphological information. In this work, we attempted to mitigate the genre differences by transforming the MSA-ATB to look like TD-ATB. This will allow creating in tandem a bilingual lexicon with different dialectal concepts (Figure1). For this purpose, we adopted a transformation method based on the parts of speech of ATB's words, as discussed in the following.

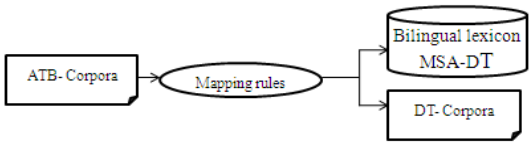

Figure1- Method for creating TD resources

4 Mapping rules based on verbal morphological distinction

There's a difference between verb conjugation in MSA and that in TD. We find that in TD, the gender distinction is not marked. Most Tunisian people do not distinguish between masculine and feminine with the second person-singular. Similarly, we mark the absence of the masculine and feminine dual. Another conjugation difference is in the passive form of the TD and MSA verb. In fact, the passive form of most Tunisian verbs is obtained by preceding the verb with the consonant 'ت' [t]. Unlike in MSA, passive verbs in TD cause the transformation of the structure of the sentence: For example, the transformation of the sentence (Active voice) كلا الطفل التفاحة/klA alTfol AltofeHap/The boy ate the

apple/ is in passive voice "التفاحة تاكلت"/AltofeHa **t**eklit/The apple has been eaten/

In the imperfect, the [t] lies between the root and the prefix as in the following:

"يتاكل"/**yi**tekil/ The lunch (M) is edible. *Masculin*

/"تتاكل"/titekil/The apple (F) is edible *Feminin.* In addition to this type of form, the dialect offers another form frequently used as question such as: "تتاكلشي"/**ti**tekil**$y**/ Is it edible?

In this work, as we aim to build a lexicon for Tunisian verbs, we must take into account these differences. But to define Tunisian inflected forms, we should first define the main concept of "Arabic verb" and we will do this by studying the morphological and lexical differences that may exist between TD verbs and MSA verbs. Indeed, in Arabic there are three principal verbal concepts:

1-Root: It is the basic source of all forms of Arabic verbs. The root is not a real word; rather, it is a sequence of three consonants that can be found in all words that are related to it. Most roots are composed of three letters; very few are composed of four or five consonants.
2-Pattern: In MSA, patterns are models with different structures that are applied to the root to create a lemma. For example, for the root خ ر ج: xrj, we can apply different patterns which give different lemmas with different meanings:
Root1: xrj/خ ر ج/ C1C2C3+ verbal pattern1: AistaC1oC2a3 =lemma1 اسْتَخْرَجَ/ to extract
Root1: xrj/خ ر ج/C1C2C3+ verbal pattern2 FoEaL (FaEal)=lemma2 خَرَجَ/ to go out .
Root1: xrj (خ ر ج)/C1C2C3+ verbal pattern3 >aC1oC2aC3=lemma3 أَخْرَجَ/ to eject
3-Lemma: The lemma is a fundamental concept in the processing of texts in at least some languages. An Arabic word can be analyzed as a root inserted into a pattern.

4.1 Verbal concepts for the Tunisian dialect

As we aim to adapt MSA tools to TD, we tried to build for TD verbs the same concepts as those in MSA. Therefore, we focused in this work on the study of correspondences that may exist among the concepts of MSA verbs and dialect verbs. First, we extracted all the verbs that exist in ATB, represented in their inflected forms. Second, we used a lemmatizer to extract lemmas; we obtained as a result 1500 different MSA lemmas. Third, we built manually lemmas corresponding to TD. Later, we tried to build verbal patterns equivalent to those in MSA. Finally, since there is no standard definition of roots in TD, we opted for a deductive method to define root for dialect verbs. Figure 2 illustrates this method.

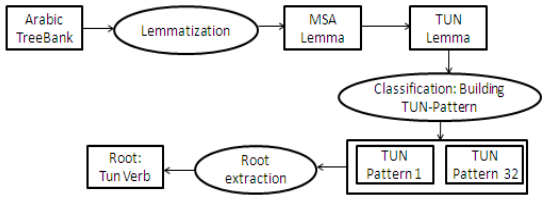

Figure 2: From ATB verb to TD-verb

Building TD-lemmas: Verbs in the ATB corpus are presented in their inflected forms. So, we extracted lemmas and their roots using the morphological analyzer developed by Elexir FM (Smrz, 2007). As we are native speakers of TD, we associated to each MSA-Lemma a TUN-Lemma. As a result, we found that 60% of the verbs change totally when passing from MSA to TD. This is a preliminary step for building Tunisian patterns from which we will be able to deduct the inflectional forms. So, as we have 1500 TD-Lemmas, and starting from the fact that MSA verbs have patterns describing their morphological behavior during conjugation, we tried, whenever possible, to define to each TD-Lemma a TD-Pattern which is similar to the MSA-pattern.

Building TD-patterns: The challenge in building TD-patterns was to find patterns similar to those in MSA. In MSA, patterns are models with different structures that are applied to the root to create a lemma. In fact, for trilateral roots there are in MSA ten patterns I: CCC, II: CaC~aC, III: CACaC, IV: >aCCaC, V: /taCaC~aC, VI: taCACaC,VII:AinCaCaC, VIII:AiCtaCaC, IX : AiCCaC~, X: AistaCCaC.
To classify the lemmas that we have already built, we focused on the creation of verbal patterns for TD verbs. So, we chose three criteria that classify verbs from general (without considering the vowels of the word) to specific (dealing with the different variations of vowels in its conjugation).

Classification according to the verb model
Verb model means the form that the root takes after applying the Pattern, for example:
Root :خ ر ج /xrj ; Pattern: CaCaC; Lemma : خَرَجَ/ xaraj ; Model : CVCVC

Root :خ ر ج /xrj ; Pattern: AistaCCaC ; Lemma: اسْتَخْرَج/Aistaxraj ; Model : AistaCVCVC
Classification according to the model of the verb consists in studying similarities between verb models without considering changes in vowels. Indeed, as we have already mentioned, we have 40% of verbs that do not change their root when the pass from MSA to TD. They therefore have the same model without considering vowels. To do this, we assigned to TD-verbs patterns equivalent to those in MSA (1).
For example: MSA-lemma: خَرَجَ / xaraj/go out
Pattern-MSA: CaCaC Model: CVCVC
→ TD: lemma: xoraj Model: CVCVC then Pattern-TUN: CoCaC
Moreover, for verbs that change their root when passing to the dialect, we reasoned as follows: For a TD verb whose model looks like the model of a TD-verb for which we have already assigned a Tun-pattern (1), we assign the same Tun-pattern (2).
Example1:
MSA: صَمَت /Samat /be silent → TD: سْكُت /sokut
Model : CVCVC looks like the model of خْرَج / xraj/: go out : CVCVC. (1)
We have already assigned to خْرَج/ xoraj the -TUN-pattern: CoCaC. Therefore, سْكُت/ sokut will have the pattern -TUN: CoCuC (2).
In this way, we classified almost all TD verbs except a few who have a complex form illustrated by a verbal unit plus another lexical unit (particle or other...).
For example, the translation of the MSA verb رافَقَ/rAfaqa/go with → is in TD: مْشَى-مع/mo$aY-moEa. We associated this type of verb to patterns that we called "exception patterns"

Classification according to the vowel of the second consonant of the pattern

The vowel of the second consonant of the pattern (vowel letter ع / E) is a fundamental criterion for classifying a verb in MSA (Ouerhani, 2009). In fact, according to this criterion, the MSA pattern I is divided into six patterns due to the variation of the vowel of the second consonant (both in past and present tense). These patterns are respectively: I-au: CaCa-yaCoCuC ; I-ai : CaCaC-yaCiC ;
I-aa: CaCaC-yaCoCaC, I-ia: CaCiC-yaCoCaC ; I-uu: CaCuC-yaCoCuC ; I-ii: CaCiC-yaCaCiC.

In TD, this variation is very common and it is marked not only in the pattern I but in all patterns. For this reason, we proposed to divide these patterns and to define new patterns in order to consolidate the verbs that have the same behavior. For example, for the Pattern-TUN II:
MSA: Pattern-TUN II: no TD sub-pattern: New three sub-patterns: II-aa: CaC~aC/yiCaC~aC ; II-ai: CaC~aC /yiCaC~iC ; TUN II-ii: CaC~iC /yiCaC~iC

Classification according to the Imperfect mark

The third classification criterion is based on the imperfect mark. In MSA, this mark remains unchanged in all verbs belonging to the same class. In fact, for the MSA pattern I CaCaL/**ya**CCAC, the mark is **ي/ya** ; for example: كَتَبَ -يَكْتُبُ/kataba-**ya**ktubu/write. For the pattern III/CACaC/yuCACiC, the mark is **يُ/yu**; for example شارَك-يُشارِكُ/$Araka-**yu**$Ariqu/participate.
However, we noticed that in TD, this regularity appears especially in the pattern I, so this mark can vary even within the same class. For example, خرج - يخرج / xraj-**yu**xruj/to go out belongs to theTUN –pattern-I-au; قال - يقول / QAl-**yi**quwl/to say belongs to the TUN-pattern-I-au. Note that although these two verbs belong to the same class, their imperfect marks are different. For this reason, we proposed to extend the TUN-pattern-I-au and define more sub-patterns for the pattern I.
In this way, we assigned to خرج - يخرج / xraj-**yu**xruj the pattern- I -au-**u** and to قال - يقول / QAl-**yi**quwl the pattern- I -au-**i**.
The result of this classification has allowed distinguishing 32 patterns for dialect verbs while there were 15 in MSA.

-<u>TD-root definition</u>

In Tunisian dialect, there is no standard definition for the root. For this reason, dialect root construction was not obvious, especially when the verb root changes completely from the MSA to the dialect. In fact, to define a root for TUN verbs, we adopted a deductive method. Indeed, in MSA, the rule says: root + pattern =Lemma (1). In our case, we have already defined the TUN-lemma and the Tun-pattern. Following rule (1), the extraction of the root is then made easy. For example, we classified the lemma إستنى / Aistan ~ aY / Wait in the pattern AistaCCaC then root(?) + AistaCCaC= Aistan~Y
Following (1), the root is "ننني" [NNY]. In fact, we can say that the definition of roots is a problematic issue which could allow more discussion. According to (1), it was as if we had forced the roots to be [NNY]. However, if we

classify إستنى / Aistann ~ aY under the pattern AiCtaCaC, the root in this case must be سنن / snn. The root can also be quadrilateral سنني / snnY if we classify Aistann ~ aY under the pattern AiCCaCaC. But as there's no standard, we did our best to be as logical as possible to define dialectal root.

4.2 Verbal lexicon structure

The various verbal transformations described above are modeled and stored in a dictionary of verbs as follows: to each MSA verbal block containing the MSA-lemma, the MSA-pattern and the MSA-root will correspond a TD- block which contains the TD-lemma, the TD-root- and the TD-pattern. So, knowing the pattern and the root, we will able to generate automatically various inflected forms of the TUN verbs. That's why we also stored in our dictionary the active and the passive form of the TD-lemma in perfective and imperfective tenses. We also stored the inflected forms in the imperative (CV). Figure 3 shows the structure that we have defined for the dictionary to present the TD-verbal concepts (in Section 4, we will explain how we will automate the enrichment of this dictionary)**.**

```
<DIC_TUN_VERBS_FORM>
  <LEXICAL-ENTRY POS="VERB">
<VERB ID-VERB="48">
    <MSA-LEMMA>
      <Headword-sa>عَايَنَ</Headword-MSA
      <Pattern>فاعل</Pattern>
      <Root-Msa>عين</Root-Msa>
      <Gloss lang= "ang" > Observe</Gloss>
    </MSA-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>
:::
</DIC_TUN_VERBS_FORM>
```

Figure3- Verbal dictionary structure

5 Mapping rules based on syntactic distinction

We identified three areas that reflect the specific syntax of the dialect: word order, grammatical negation and syntactic tools categories. In the following section, we will explain how we define these dialect structures in our lexicon.

5.1 Word order

The order of the elements in the dialect sentence seems to be relatively less important than in other languages . However, the canonical word order in Tunisian verbal sentences is SVO (Subject-Verb-Object) (Baccouche, 2003).
In contrast, the MSA word order can have the following three forms: SVO / VSO / VOS (2).
(1) TD: « الطْفُلْ كْتِبْ الدَرْسْ ».SVO
(2) MSA: « كتب الطفل الدرس ».VSO.
This opposition between MSA and the dialect is clearer in the case of proper names. In fact, MSA order is VSO (3) while the order in TD is SVO. **(**Mahfoudhi, 2002)
(3) MSA : « ضرب موسى عيسى ».
(4) TD : « موسى ضْرَبْ عيسى ».
There are other types of simple dialect sentences named nominal sentences which do not contain a verb. They have the same order in both TD and MSA. For example:
MSA: الطقس حار/TaKs HAr/ The weather is hot

TD: الطَقْسْ سْخُونْ /TaKs sxuwn/The weather is hot

In our work, we discussed some nominal groups at the syntactic level. The word order is generally reversed when passing to TD. For example
(1) MSA: ADV + ADJ
أيضا >ayDaA/Also+مْثَقَّف /muvaK~af/also educated
TD: ADJ +ADV
ADJ/ مْثَقَّف +ADV/زاده
(2) MSA: Noun + ADJ
MSA: كُتُب كَثِيرة /kutubun kavira/many books->
TD: ADJ + Noun
TD: برشا كتُب /bar$A ktub
In the dictionary, we present this kind of rule as shown in Figure 4.

```
<ADV-MSA ID="5">
    <MSA-LEMMA>أَيْضاً</MSA- LEMMA>
    <GLOSS ang="ang">Also</GLOSS>
    <CONTEXT ID="1">
      <CONFIG ID="1" Position="Before" POS="ADJ" />
        <TOKEN>
          <TUN ID="1" DIC="ADJECTIVES" POS="ADJ" />
          <TUN ID="2" />
         <TUN ID="3">زَادَا</TUN>
        </TOKEN>
```

</CONTEXT>

Figure 4- Syntactic rule representation in the dictionary

5.2 Grammatical negation

Negation particles are generally set before the verb and can sometimes change the combination. For example, if the word أكتب> />ktb /Write in MSA is preceded by a negative particle such as لم/lam (Do not), the verb in the dialect will be: ماكتبتش/mAktibti$=
TUN-Neg-Particle(ما)+ Tun-verb (كتب)+ Tun-Neg-enc (ش)

5.3 Syntactic tool categories

Tools words or Syntactic tools exist in a large amount in the Treebank and all MSA-texts. However, their transformation was not trivial and required for each tool a study of its different contexts.
A tool word may have different translations depending on its context. For example, the particle حَتَّى/ HatY/so that: we found this particle in ATB in three contexts. This particle gives a new translation whenever it changes context:
1- حَتَّى/ HatY + verb = باش (TUN-particle) + TUN_verb
2 حَتَّى/ HatY + NEG_PART = باش (TUN-particle) + TUN_NEG_PART
otherwise
3-حَتَّى /HatY =حَتَّى /HatY
So, to deal with these transformations, we converted them into rules and stored them into a lexicon of tool word transformation.

Context dependent transformation
We mean by context dependent transformation the passage MSA-TD which is based on transformation rules. Indeed, given the word MK, we say that the transformation of MK is based on context if it gives a new translation whenever it changes context. RTK : X + M + **Y = TDk**

$X = \sum_{j=1}^{m} Mj : POSj$; $Y = \sum_{i=1}^{n} Mi : POSi$; k varies from 1 to z ;

RTk: transformation rules n°k; POS : Part of speech ; M: word tool, TDk: Translation n°k

The transformation of a tool word may depend on the words (X) that precede it, or on the following word (Y), or both. If none of the contexts is presented, then a default translation will be assigned to the tool word. In total, we defined in the tool words dictionary 316 rules for the 146 ATB's tool words.

In the dictionary, we presented a transformation rule. In fact, for each tool word we defined a set of contexts; each context contains one or more configurations. The configuration describes the position and the part of speech of the words of context. Each context corresponds to a new translation of the tool word (Figure 5).

```
<PREP-MSA ID="9">
   <MSA-LEMMA>حَتَّى</MSA-LEMMA>
   <GLOSS lang="ANG ">until </GLOSS>
    <CONTEXT ID="1">
     <CONFIG ID= "1" Position="Before" PRC="DET" />
                  <CONFIG ID="2" Position=" Before "
                    POS="NOUN">ساعَة</CONFIG>
<CONFIG ID="3" Position=" Before" POS="NOUN_NUM" />
      <TOKEN>
       <TUN ID="1">حَتَّى-ل</TUN>
       <TUN ID="2" POS="NOUN_NUM" />
      </TOKEN>
     </CONTEXT>
     ……
     <CONTEXT ID="6">
       …..
</Prep-MSA>
```

Figure5- Structure of a context dependent rule in the dictionary

Context independent transformation

In addition to the context-dependent transformations, the translation of some tool words in the corpus was direct "word to word"; the word remained the same regardless of the context. Figure 6 shows an example of how we represented this kind of translation in the dictionary

```
<SUB_CONJ-MSA ID="7">
<MSA-LEMMA>كَيْ</MSA-LEMMA>
<GLOSS lang="ANG">In order to
</GLOSS>
<TOKEN>
<TUN ID="1">بَاشْ</TUN>
</TOKEN>
</SUB_CONJ-MSA>
```

Figure 6- Structure of a context independent rule in the dictionary

6 Automatic generation of Tunisian Dialect corpora

To test and improve the developed bilingual models, we exploited our dictionaries to

automate the task of converting MSA corpora to corpora with a dialect appearance.

For this purpose, we developed a tool called Tunisian Dialect Translator (TDT) which enables to produce TD texts and to enrich the MSA-TD dictionary (Figure 6). The TDT tool works according to the following steps:

1-Morphosyntactic annotation of MSA texts: TDT annotates each MSA text morpho-syntactically by using the MADA analyzer (Morphological Analyser and disambiguisator of Arabic) (Habash, 2010). MADA is a toolkit which, given a raw MSA text, adds lexical and morphological information. It disambiguisates in one operation part-of-speech tags, lexemes, diacritizations and full morphological analyses.

2-Exploiting MSA-TD dictionaries: Based on each part of speech of the MSA-word, TDT proposes for each MSA structure the corresponding TD translation by exploiting the MSA-TD dictionaries.

3-Enriching the lexicon: As our MSA-TD dictionaries do not cover all Arabic words, texts resulting from the previous step are not totally translated. Therefore, in order to improve the quality of translation and to enrich our dictionaries, enabling them to be well used even in other NLP applications, we added to TDT a semi-automatic enrichment module. This module filters first all MSA words for which a translation has not been provided. Then, TDT assigns to them their corresponding MSA-lemmas and POS. If the POS is a verb or a noun, the user proposes a TD-root and a TD-pattern (described in subsection 3.2) and the TDT generates automatically the appropriate Tunisian lemma and its inflected forms.

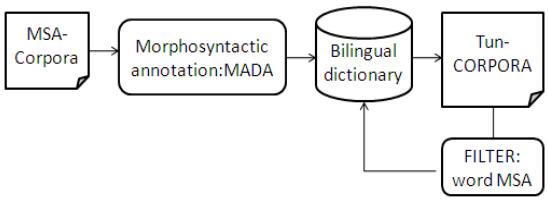

Figure7- Automatic process for generating Tunisian corpora

7 Evaluation

To evaluate translations of verbs occurring in our MSA-TD dictionary, we asked 47 judges (native speakers) to translate for us a sample containing 10% of verbs extracted from the dictionary. In this sample, there are 52 verbs that don't change their root when passing to TD and 98 do otherwise. The evaluation consists in comparing what we have proposed as a translation of lexical items taken from the ATB with the proposals of judges who are native speakers of the Tunisian dialect. The percentage of agreement between the judges' translations and the translations proposed in our lexicon is calculated. Table 1 shows the results obtained

Verbs	Unchanged	Changed	Total
Number of verbs in the sample	52	98	150
Agreement	97,17%	63,21%	74,97%

Table 1- Evaluation of verb translation

Moreover, as the translation of the majority of tool words depends on context, we asked 5 judges to translate 89 sentences containing 133 tool words. In this sample, we repeated some tool words in the same sentence but in a different context. Table (2) gives the percentages of agreement between the translations of the judges and those in our dictionaries of tools words. The variation in percentage is due to the fact that for some words, the judges do not agree among themselves. The table shows also the percentage of disagreement between the judges and the dictionaries.

	2 juges	**3 juges**	**4 juges**	**5 juges**
Agreement	72,69 %	74,53 %	71,34 %	71,23 %
Disgreement	18,79 %	15,03 %	14,28 %.	12,03 %

Table 2- Evaluation of tool word translation

In fact, disagreement arises when no judge gives a translation similar to the translation proposed in the dictionaries. But, by increasing the number of judges, the disagreement decreases, which proves that our dictionaries contain translations accepted by several judges

8 Conclusion

This paper presented an effort to create resources and translation tools for the Tunisian dialect.

To deal with the total lack of written resources in the Tunisian dialect, we described first a method that allowed the creation of bilingual dictionaries with in tandem TD-ATB. In fact, TD-ATB will serve as a source of insight on the phenomena that need to be addressed and as corpora to train TD-NLP tools. The verb dictionaries and the verbal concepts that we have developed were also exploited in order to adapt MAGEAD

(Habash *et al.* 2006) (Morphological Analyser and Generator of Arabic Dialect) to the Tunisian dialect (Hamdi *et al.*, 2013).

We focused second on describing TDT, a tool used to generate automatically TD corpora and to enrich semi-automatically the dictionaries we have built.

We plan to continue working on improving the TD-resources by studying the transformation of nouns. We also plan to validate our approach by measuring the ability of a language model, built on a corpus translated by our TDT tool, to model transcriptions of Tunisian broadcast news.

Experiments in progress show that the integration of translated data improves lexical coverage and the perplexity of language models significantly.

References

Al-Sabbagh Rania and Girju Roxana. 2010. *Mining theWeb for the Induction of a Dialectical Arabic Lexicon.*In Nicoletta Calzolari.

Bies Ann. 2002. *Developing an Arabic Treebank: Methods , Guidelines , Procedures , and Tools*.

Baccouche Tayeb. 1994. *L'emprunt en arabe moderne*, Beit Elhikma et IBLV, Tunis.

Baccouche Tayeb. 2003. *La langue arabe: Spécificités et évolution*.

Brustad Kristen. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Chiang David, Diab Mona, Habash Nizar, Rambow Owen and Shareef Safiullah. 2006. *Parsing Arabic Dialects*. In Proceedings of the European Association for Computational Linguistics (EACL).

Chiang David, Diab Mona, Habash Nizar, Rambow Owen and Safiullah Shareef. 2006. *Parsing Arabic Dialects*. In Proceedings of the European Chapter of ACL. EACL.

Diab Mona, Habash Nizar, Owen Rambow, Al Tantawy Mohamed and Benajiba Yassine. 2010. *COLABA: Arabic Dialect Annotation and Processing*. LREC Workshop on Semitic Language Processing, Malta, May 2010.

Graja Marwa, Jaoua Maher and Belguith Lamia. 2011. *Building ontologies to understand spoken*, CoRR.

Habash Nizar, Rambow Owen and Roth Ryan. 2009. *MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization.* In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Habash Nizar and Rambow Owen. 2005. *Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop.* InProceedings of the 43rd Annual Meeting of the Association for Computational Linguistics ACL'05, pages 573–580, Ann Arbor, Michigan.

Habash Nizar and Rambow Owen. 2006. *Magead: A morphological analyzer for Arabic and its dialects*. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL'06), Sydney, Australia.

Hamdi Ahmed, Boujelbane Rahma, Habash Nizar and Nasr Nizar. 2013. *Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde*. TALN, Nante, France.

Hitham Abo Bakr, Shaalan Khaled and Ibrahim Ziedan. 2008. *A hybrid approach for converting written egyptian colloquial dialect into diacritized Arabic*. In the 6th International Conference on Informaticsand Systems, INFOS. Cairo University.

Holes Clive. 2004. *Modern Arabic: Structures, Functions, and Varieties. Georgetown Classics in Arabic Language and Linguistics*. Georgetown University Press.

Marcel Diki-kidiri. 2007. *Comment assurer la présence d 'une langue dans le cyberespace*.

Maamouri Mahmoud and Bies Ann. 2004. *Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools*, Workshop on Computational Approaches to Arabic Script-based Languages, COLING.

Maamouri Mahmoud, Bies Ann, Krouna Sondes, Kulick Seth, Mekki Wigdan and Buckwalter Tim. 2009. *Penn arabic treebank guidelines with much appreciated contributions from*, 1–248.

Maamouri Mohamed, Bies Ann, Kulick Seth, Zaghouani Wajdi, Graff David and Ciul Michael. 2010. *From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News*, Lrec.

Mohamed Emad, Mohit Behrang and Oflazer Kemal. 2012. Transforming Standard Arabic to Colloquial Arabic, (July), 176–180.

Mahfoudh Abdessatar. 2002. *Agreement lost Agreement Regained: A Minimalist Account of Word Order and Agreement Variation in Arabic*, University of Ottawa.

Nimaan Abdillahi, Nocera Pascal and Orres-Moreno Juan-Manuel. 2006. *Boîte à outils TAL pour des*

langues peu informatisées: le cas du Somali, JADT.

Ouerhani Bechir, *Interférence entre le dialectal et le littéral en Tunisie: Le cas de la morphologie verbale*, 75–84.

Scherrer Yyves. 2008. *Transducteurs à fenêtre glissante pour l'induction lexicale*, Genève

Seng Sopheap, Sam Sethserey, Le Viet-Bac, Bigi Brigitte and Besacier Laurent. 2010. *Reconnaissance automatique de la parole en langue khmère : quelles unités pour la modélisation du langage et la modélisation acoustique*.

Smrž Otakar. 2007. *Computational Approaches to Semitic Languages*, ACL, Prague

Smrž Otakar, Viktor Bielický, Iveta Kourilová, Jakub Kráčmar, Jan Hajic and Petr Zemánek. 2008. *Prague Arabic Dependency Treebank: A Word on the Million Words*.

Zribi Ines, Graja Marwa, Ellouze Khmekhem Mariem, Jaoua Maher and Hadrich Belguith Lamia. 2013. *Orthographic Transcription for Spoken Tunisian Arabic*, CICLing, Samos, Greece.