Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model

Rahma Boujelbane
Miracl Laboratory, ANLP Research Group, University of Sfax, Tunisia
Rahma.boujelbane@gmail.com

Mariem Ellouze khemekhem
Miracl Laboratory, ANLP Research Group, University of Sfax, Tunisia
mariem.ellouze@planet.com

Siwar BenAyed
Faculty of Economics and Management of Sfax
siwar.ben.ayed@gmail.com

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

Abstract

Since the Tunisian revolution, Tunisian Dialect (TD) used in daily life, has became progressively used and represented in interviews, news and debate programs instead of Modern Standard Arabic (MSA). This situation has important negative consequences for natural language processing (NLP): since the spoken dialects are not officially written and do not have standard orthography, it is very costly to obtain adequate corpora to use for training NLP tools. Furthermore, there are almost no parallel corpora involving TD and MSA. In this paper, we describe the creation of Tunisian dialect text corpus as well as a method for building a bilingual dictionary, in order to create language model for speech recognition system for the Tunisian Broadcast News. So, we use explicit knowledge about the relation between TD and MSA.

1 Introduction

Recently, due to the political changes that have occurred in the Arab world, we noticed a new remarkable diversity in the media. Arabic dialects used in daily life have became 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. Since that, the media programs have been changed: television channels, political debates and broadcasts news have been multiplied. Therefore, this gave birth to a new kind of language. Indeed, the majority of speech is no longer on MSA but alternating between MSA and dialect. Thus, we can distinguish in the same speech, MSA words, TD words and MSA-TD words such as a word with an MSA component (stem) and dialectal affixes. This situation poses significant challenges to NLP, in fact applying NLP tools designed for MSA directly to TD yields significantly lower performance, making it imperative to direct the research to building resources and tools to process this kind of language. In our case we aim to convert this new language to text, but 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-MSA resources, whether parallel text or paper dictionaries. In this paper, we describe a method to create Tunisian Dialect (TD) text corpora and the associated lexical resources as well as building bilingual dictionary MSA-TD.

2 Related work

Spoken languages which have no written form can be classified as limited-resources languages. Therefore, several studies has attempted to overcome the problems of computerization of these languages. (Scherrer, 2008) in order to computerize the existing dialect in Switzerland, developed a translation system: standard German to any variety of the dialect continuum of German-speaking Switzerland. Moreover, (Shaalan et al, 2007) proposed a system of translation MSA-Egyptian dialect. For this, they tried to build a parallel corpus between Egyptian dialect and MSA-based on mapping rules EGY-MSA. Besides dialects, there are several languages from the group of limited-resources 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

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 88–93,
Sofia, Bulgaria, August 8, 2013.

process the Somali language: Collecting corpus from the web, automatic synthesis of texts and machine translation French-Somali. (SENG, 2010) selected news sites in Khmer to collect data in order to solicit the lack of resources in Khmer.

The literature shows that there is little work that dealt with the Tunisian Arabic, the target language of this work. (Graja et al, 2011), for example, treated the Tunisian Dialect for understanding speech. To train their system, researchers relied on manual transcripts of conversations between agents at the train station and travelers. However, a limited vocabulary is a problem if we want to model a language model for a system of recognition of television's programs with a wide and varied vocabulary.

3 Method to create Tunisian Dialect Corpora

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

3.1 Penn Arabic TreeBank corpora to create bilingual lexicon MSA-TD

Treebanks, are an important resources that allows 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 and Haji, 2007; Smrž et al., 2008). The PATB not only provides tokenization, complex POS tags, and syntactic structure; it also provides empty categories, diacritizations, lemma choices. The PATB consists of 23,611 parse-annotated sentences (Bies and Maamouri, 2003; Maamouri and Bies, 2004) from Arabic newswire text in MSA. The PATB 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 us to create in tandem a bilingual lexicon with different dialectal concept (Figure1). For this, we adopted a transformation method based on the parts of speech of ATB's word.

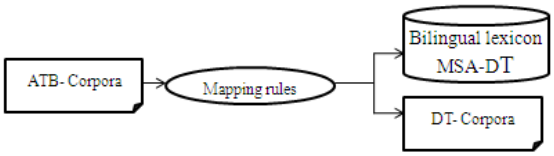

Figure1- Methodology for creating TD resources

3.2 Modeling verbal lexical entries for the bilingual dictionary

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 correspondence that may exist among the concepts of MSA verbs and dialect verbs. In Arabic there are three principal verbal concepts:

1-Root: It is the basic source of all forms of Arabic verb. 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 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 اِسْتَخْرَجَ/ to extract*

Root1*: xrj/خرج/C1C2C3+ verbal pattern2 FoEaL(FaEal)=lemma2 خَرَجَ/ to go out .*

Root1: *xrj (خ ر ج)/C1C2C3+ verbal pattern3 >aC1oC2aC3=lemma3 أَخْرَجَ/ to eject*

2-Lemma: The lemma is a fundamental concept in the processing of texts in at least some languages. Arabic words can be analyzed as consisting of a root inserted into a pattern.

TD-lemma building: Verbs in the PATB 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 associate to each MSA-Lemma a TUN-Lemma. As a result, we found that 60% of verbs change totally by passing from MSA to TD. As we have 1500 TD-Lemmas, and starting from the fact that

MSA verbs have patterns describing their morphological behavior during conjugation, we tried to assign, if possible, to each TD-Lemma a TD-Pattern.

TD-pattern building: The challenge on building TD-pattern was to find patterns similar to those in MSA. Thus, by studying the morphology of TD-lemmas, we remarked that it's possible to assign to TD-lemmas the same pattern as those on MSA but with defining other patterns that will be sub-patterns to these patterns. In fact, this process has allowed distinguishing 32 patterns for dialect verbs while there were 15 in MSA. This was due to the morphological richness and the frequent change of vowel within TD-lemmas. For example:

In MSA *$Arak/yu$Arik/to participate* and *dAfaE/yudAFiE/to defend* belongs to the *patternII: CACaC(perfectiveform)/yiCACiC* (imperfectiveform). In TD the model of these tow verbs remains *CACVC/yVCACVC* but the vowel of the second consonant of the pattern (vowel letter ع / E) change. The mark of this vowel is a fundamental criterion for classifying a verb in MSA (Ouerhani, 2009), that's why we proposed to define tow sub-pattern for the pattern II, by dividing the pattern-II to *II-i: CACiC/yVCACiC* and *II-a:CACaC/yVCACaC*. As consequence, *$Arak/yu$Arik/* becomes in TD *$Arik/yi$Arik/* belongs to *CACiC/yVCACiC* and *dAfaE/yudAFiE* becomes in TD *dAfaE/yidAFaE* belongs to *CACaC/yiVCACaC.*

Therefore, by adopting this reasoning, we succeeded with the ATB's verbs to define pattern for the TD verb. Thus, knowing these new patterns, we will be able to assign a pattern for all TD verbs.

TD-root building: In Tunisian dialect, there is no standard definition for the root. For this, construction of root dialect was not obvious, especially when the root verb changes completely through the MSA to the dialect. In fact, to define a root for TD verbs, we have adopted a deductive method. Indeed, in MSA, the rule says: root + pattern= Lemma (1). In our case, we have already defined the TD-lemma and the TD-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= إستنى / ~ YAistana~*

Following (1), the root for the verb *إستنى /Aistan ~ aY/Wait* is *"ني" [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 like we have forced the roots to be *[NNY]*. However, if we classified *Aistann ~ aY* under the pattern *AiCCaCal*, the root in this case must be *snn*. The root can also be quadrilateral سنني / snnY if we classified *Aistann~ aY* under the pattern *AiCCaCaC*. But as there's no standard, we have done in our best to be the most logical possible to define dialectal root.

3.3 Structure of verbal lexicon entries

Different verbal transformations described above are modeled and stored at a dictionary of verb as follows: to each MSA verbal block containing MSA-lemma, MSA-pattern and MSA-root will correspond TD- block which containing TD-lemma, TD-root and 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 stored in our dictionary the active and the passive form of the TD-lemma in perfective and imperfective tense. We also store the inflected forms in the imperative (CV). Figure 2 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= "fr" > Observer</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>
```

Figure2- Verbal structure in dictionary

3.4 Modeling lexical entries for tools words in the bilingual dictionary

Tools words or syntactic tools are an area that reflects the specific syntax of the dialect. It has 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 context. In our approach, we defined two kinds of transformations. The first requires the study of different context of a tool word. In fact, the same word may have different translations depending on its context. Thus, to deal with the variation of context, we developed mapping rules. Note that among these contexts, there are those that cause a change in the syntactic order of words by passing to the dialect. The second transformation is direct, the word remains unchanged whatever the context.

3.5 Context dependent transformation

We mean by transformation-based context, the passage MSA-DT which is based on transformation rules. Indeed given a word W, we say that the transformation of W is based on context if it gives a new translation whenever it changes on context. RT : X + W + Y = TDk

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

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

The transformation of a tool word may depend to the words that it precedes (X), or the following word (Y), or both. If none of the contexts is presented, then a default translation will be assigned to the word tool. For example, For the tool word *"حتى" [hatY]/So that* which have the POS: Preposition, we developed three different mapping rules depending to the context in the ATB corpora.

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

In total, we developed 316 rules for the ATB's tools words. Figure 3 shows how we present a transformation rule in the dictionary. For each tool word we have 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.

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

Figure3- Context dependent rule structure in dictionary

Syntactic transformation:

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 , 2004). In contrast, MSA word order can have the following three forms: SVO / VSO / VOS (2).

(1) TD: الطِّفْل كْتِبْ الدَرْسْ /AlTfol ktib aldars/the child wrote the lesson: SVO

(2) MSA: كتب الطفل الدرس/ktib Altfol Ald~ars/wrote the boy the lesson: VSO.

This opposition bestween the 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: أكل القط الفئران />akal Alqit Alfi>rAn / Cats rats

(4) TD: القط أكل الفئران / Alqit >akal Alfi>rAn /Cats eat rats

There are other types of simple dialect sentences named nominal sentences which do not contain a verb. They have the same order in both Tunisian and MSA. For example:

MSA: حار الطقس /TaKs HAr/ weather is hot
TD: سْخُونْ الطَّقْسْ / TaKs sxuwn/ weather is hot

In our work, we discussed the syntactic level at some nominal groups. The word order is generally reversed by passing to TD. For example:

(1)MSA: ADV + ADJ:
أيضا />ayDaA/Also+ مْثَقَّف /muvaK~af/also educated
(2) TD: ADJ +ADV:
TD: ADJ/ مْثَقَّف +ADV/ زاده
(2)MSA: Noun + ADJ:
كُتُب كَثِيرة /kutubun kavira/many books

TD: ADJ + Noun:
برشا كتُب /bar$A ktub
In the dictionary, we present this kind of rule as shown in the figure 4.

```
<ADV-MSA ID="5">
<MSA-LEMMA> ا أَيْض </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=" زَادَا<" 3 </TUN>
</TOKEN>
</CONTEXT>
```

Figure 4- Syntactic rule representation in the dictionary

3.6 Context independent transformation

In addition to the context-dependent transformations, the translation of some tools words in the corpus was direct "word to word", eg; the word remains the same regardless of the context. Figure 5 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 5- Direct translation structure in the dictionary

4 Automatic generation of Tunisian Dialect corpora

To test and improve the developed bilingual models, we tried by exploiting our dictionaries to automate the task of converting MSA corpora to a corpora with a dialect appearance.

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

1-Morphosyntactic annotation of MSA texts: TDT annotate each MSA text morphosyntactically by using MADA analyzer (Morphological Analyser and disambiguisator of Arabic) (Habash, 2010). MADA is a toolkit that, given a raw MSA text, adds as much lexical and morphological information as possible by disambiguating 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 propose for each MSA structure the corresponding TD translation by exploiting the MSA-TD dictionaries.

3-Enriching lexicon: As the lexical database does 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 to be well used even in other NLP application, 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 assigned for them their corresponding MSA-lemmas and POS, the user proposes, if the POS is verb or noun, a TD-root and a TD-pattern (described in subsection 3.2) and the TDT proposes automatically the appropriate Tunisian lemma and it's inflected forms.

5 Evaluation

To evaluate different translations of the verbs dictionary, we asked 47 judges (native speakers) to translate a sample containing 10% of verbs in the dictionary. 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 Tunisian dialect. The percentages calculated reflect the percentage of agreement for each verb translations between judges and the translation proposed in our lexicon. Table 1 shows the obtained results.

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

For the same context, an MSA-Verb may have many translations. The agreement decreases for changed verbs because the judges may propose a valid translation different from what we have proposed in the dictionary. Moreover, as the translation of the majority of tool words depends on context, we asked 5 judges to translate 89 sentences containing 133 words tools. In this sample, we made some tools words repeated in the same sentence but in different context. Table

(2) gives the percentages of agreement between the translations of the judges and those of our dictionaries of tools words. The variation in percentage is due to the fact that for some words, judges do not agree among themselves. The table also shows the percentage of disagreement between judges and 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, the disagreement arises when no judge gives translation similar to the translation proposed in the dictionaries. But, by increasing the number of judges, the disagreement decreases which proves that our dictionaries are able to give acceptable translations by several judges

6 Conclusion

This paper presented an effort to create resources and translation tool for Tunisian dialect.
To deal with the total lack of written resource in Tunisian dialect, we described first a methodology 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. We focused second on describing TDT a tool 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 showed that the integration of translated data improves significantly lexical coverage and perplexity of language models.

References

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

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

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

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

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

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*.

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

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

Abdillahi Nimaan, Pascal Nocera, Juan-Manuel orres-Moreno. 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

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

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