Automatically building a Tunisian Lexicon for Deverbal Nouns

Ahmed Hamdi Núria Gala Alexis Nasr
Laboratoire d'Informatique Fondamentale de Marseille, Aix-Marseille Université
{ahmed.hamdi,nuria.gala,alexis.nasr}@lif.univ-mrs.fr

Abstract

The sociolinguistic situation in Arabic countries is characterized by diglossia (Ferguson, 1959) : whereas one variant Modern Standard Arabic (MSA) is highly codified and mainly used for written communication, other variants coexist in regular everyday's situations (dialects). Similarly, while a number of resources and tools exist for MSA (lexica, annotated corpora, taggers, parsers ...), very few are available for the development of dialectal Natural Language Processing tools. Taking advantage of the closeness of MSA and its dialects, one way to solve the problem of the lack of resources for dialects consists in exploiting available MSA resources and NLP tools in order to adapt them to process dialects. This paper adopts this general framework: we propose a method to build a lexicon of deverbal nouns for Tunisian (TUN) using MSA tools and resources as starting material.

1 Introduction

The Arabic language presents both a standard written form and a number of spoken variants (dialects). While dialects differ from one country to another, sometimes even within the same country, the written variety (Modern Standard Arabic, MSA), is the same for all the Arabic countries. Similarly, MSA is highly codified, and used mainly for written communication and formal spoken situations (news, political debates). Spoken varieties are used in informal daily discussions and in informal written communication on the web (social networks, blogs and forums). Such unstandardized varieties differ from MSA with respect to phonology, morphology, syntax and the lexicon. Linguistic resources (lexica, corpora) and natural language processing (NLP) tools for such dialects (parsers) are very rare.

Different approaches are discussed in the litterature to cope with Arabic dialects processing. A general solution is to build specific resources and tools. For example, (Maamouri et al., 2004) created a Levantine annotated corpus (oral transcriptions) for speech recognition research. (Habash et al., 2005; Habash and Rambow, 2006) proposed a system including a morphological analyzer and a generator for Arabic dialects (MAGEAD) used for MSA and Levantine Arabic. (Habash et al., 2012) also built a morphological analyzer for Egyptian Arabic that extends an existing resource, the Egyptian Colloquial Arabic Lexicon. Other approaches take advantage of the special relation (closeness) that exists betweeen MSA and dialects in order to adapt MSA resources and tools to dialects. To name a few, (Chiang et al., 2006) used MSA treebanks to parse Levantine Arabic. (Sawaf, 2010) presented a translation system for handling dialectal Arabic, using an algorithm to normalize spontaneous and dialectal Arabic into MSA. (Salloum and Habash, 2013) developped a translation system pivoting through MSA from some Arabic dialects (Levantine, Egyptian, Iraqi, and Gulf Arabic) to English. (Hamdi et al., 2013) proposed a translation system between Tunisian (TUN) and MSA verbs using an analyser and a generator for both variants.

Yet if the first kind of approach is more linguistically accurate because it takes into account specificities of each dialect, building resources from scratch is costly and extremely time consuming. In this paper we will thus adopt the second approach: we will present a method to automatically build a lexicon for Tunisian deverbal nouns by exploiting available MSA resources as well as an existing MSA-TUN lexicon

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 95–102,
Dublin, Ireland, August 23 2014.

for verbs (Boujelbane et al., 2013). We will use a root lexicon to generate possible deverbal nouns which will be later filtered through a large MSA lexicon.

This work is part of a larger project that aims at 'translating' TUN to an approximative form of MSA in order to use MSA NLP tools on the output of this translation process. The final lexicon for TUN deverbal nouns will be integrated into a morphological and syntactic parser for TUN.

The paper is organized as follows: in section 2 we describe and compare some morphological aspects of MSA and TUN, focusing on derivation. We then discuss in section 3 our approach to build a TUN lexicon for deverbal nouns from an existing MSA-TUN resource for verbs. Section 4 presents an evaluation of the results obtained and section 5 proposes some solutions to increase the coverage of the lexicon.

2 Arabic Morphology

Arabic words are built following two kinds of morphological operations: templatic and affixational. Functionally, both operations are used inflectionally or derivationally (Habash, 2010). In templatic morphology, a root and a pattern combine to form a word stem. A root is a sequence of three, four or five letters that defines an abstract notion while a pattern is a vocalized template which marks where the root radicals are inserted. To give an example, by combining the root ف ت ح *f t H*[1] with the verbal patterns 1a2a3 and ta1a22a3, two verbs are generated : (1) فتح *fataH* 'to open' and (2) تفتح *tafattaH* 'to bloom'. Derivation consists in replacing each digit of the pattern by the corresponding letter in the root.

Arabic verbs have ten basic triliteral patterns, which are conventionally noted with the Latin numbers I, ..., X. and two basic quadriliteral patterns (XI, XII) (Habash, 2010). A verb is the combination of a root and a pattern.

Many deverbal nouns can be derived from verbs. Nine kind of deverbal nouns (1, 2, 3 ... 9) are defined in Arabic (Al-Ghulayaini, 2010), each of them corresponds to a semantic relationship between the verb and the deverbal noun (see table 1).

1	اسم الفاعل	active participle
2	اسم المفعول	passive participle
3	مصدر	infinitive form
4	اسم المكان	noun of place
5	اسم الزمان	noun of time
6	اسم الآلة	noun of instrument
7	صفة مشبهة	analogous adjective
8	اسم التفضيل	comparative adjective
9	صيغة المبالغة	exaggerate form

Table 1: Arabic deverbal nouns

From the verbs (1) فتح *fataH* 'to open' and (2) تفتح *tafattaH* 'to bloom', many deverbal nouns can be derived, such as, فاتح *fAtiH* 'opener', مفتوح *maftuwH* 'opened' from (1), and متفتّح *mutfattiH* 'blooming', متفتّح *mutfattaH* 'blown' from (2). These deverbal nouns represent the active and the passive participles of these verbs. They are derived from the same root as the verb, using deverbal patterns which depend on the verbal pattern. Table 2 shows TUN and MSA patterns of the active and the passive participles for the first three verbal patterns.

Table 2 is just a sample of a larger table of deverbal nouns (henceforth called TUN-MSA deverbal table) that defines for every verbal pattern all deverbals which are derived from it in MSA and TUN.

[1]Arabic orthographic transliteration is presented in the Habash-Soudi-Buckwalter HSB scheme (Habash et al., 2007): (in alphabetical order)

ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي

A b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y

and the additional letters: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ىء , ħ ة, ý ى.

Verbal pattern	Deverbal noun	MSA patterns	TUN patterns
I	1	1A2i3	1A2i3, 1A2a3
	2	ma12uw3	ma12uw3
II	1	mu1a22i3	m1a22i3, m1a22a3
	2	mu1a22a3	m1a22a3, mit1a22i3
III	1	mu1A2i3	mfA2i3, m1A2a3
	2	mu1A2a3	mfA2a3, mit1A2a3

Table 2: TUN-MSA Deverbal Table

This table has been created by a Tunisian native speaker. Unlike MSA, which defines a unique pattern for each participle with all verbal patterns, table 2 shows that TUN has often more than one pattern for participles. However, for some other cases, such as the infinitive forms and nouns of instruments, MSA defines several nominal patterns. The choice of the nominal pattern depends on the verbal pattern.

The Arabic nominal derivation system is not systematic and depends on the meaning of the verbs. In fact, for semantic reasons, most Arabic verbs cannot derive all deverbal nouns. The verb فتح *fataH* 'open', for example, cannot produce the noun of place and time. However, فتح *fataH* derives the active and the passive participles فاتح *fAtiH* 'opener' and مفتوح *maftuwH* 'opened', the noun of instrument مفتاح *miftAH* 'key' and an exaggerate form فتاح *fattAH* 'conqueror'...

3 Overview of the Method

Our method consists in generating TUN and MSA pairs of deverbal nouns simultaneously: in a first step, we use the TUN-MSA deverbal table and an existing MSA-TUN dictionary of verbs in order to generate candidate pairs of deverbal nouns ($NOUN_{MSA}$, $NOUN_{TUN}$). These candidates are then filtered on the MSA side using an available MSA resource.

3.1 Generating pairs of deverbal nouns

As shown in the TUN-MSA deverbal table (Table 2), every verbal pattern in MSA produces several patterns of deverbal nouns (i.e., pattern IX[2] yields for example the infinitive form Ai12i3A3). The same applies to TUN (i.e., pattern IX yields the infinitive form 12uw3iyy). A total of 54 MSA and 52 TUN nominal patterns were defined. To generate deverbal lexicon we have used an existing TUN-MSA lexicon (Boujelbane et al., 2013) of 1500 verbs composed of pairs of the form (P_{MSA}, P_{TUN}) where P_{MSA} and P_{TUN} are themselves pairs made of a root and a verbal pattern. The TUN side contains 920 distinct pairs and the MSA side 1,478 distinct pairs. This difference shows that MSA is lexically richer than TUN. For every pair (a pattern and a root) we combined the root with all the nominal patterns corresponding to the verbal pattern on both sides (MSA and TUN) as shown in figure 1.

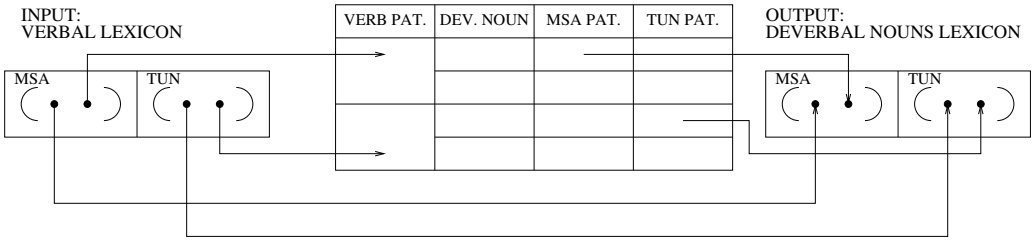

Figure 1: Generating TUN-MSA pairs of deverbal nouns using verbs

At this point, about twenty morphological and orthographic rules manually predefined are applied on the generated form in order to produce a lemma. For instance, the second root radical /y/ and /w/ changes to /ŷ/ for MSA active participle, while the second root radical /w/ changes to /y/ in the TUN side. Another

[2]The MSA and TUN IX patterns are respectively Ai12a33 and 12A3

rule which is common for MSA and TUN requires that the /t/ of the verbal pattern Ai1**t**a2a3 (VIII) and all nominal forms which derive from it, change to a /T/ if the first letter on the radical is /S/, /T/, /D/ or /Z/ : e.g. masdar اضتراب *AiDtirAb* becomes اضطراب *AiDTirAb* 'trouble'.

Following this step, a lexicon of $137,199$ nominal entries $(Noun_{MSA}, Noun_{TUN})$ was obtained.

3.2 Filtering

As it was expected, the generation method described above overgenerates: it can produce correct pairs as well as wrong pairs. Four cases have been identified:

1. Both TUN and MSA nouns are correct
2. TUN noun is wrong and MSA noun is correct
3. MSA noun is wrong and TUN noun is correct
4. Both forms are wrong

To give an example from the verbal lexicon entry (فتح, حلّ) $(fataH_{MSA}, Hall_{TUN})$ '*to open*', we can generate these four situations :

1. passive participle : (مفتوح, محلول) $(maftuwH_{MSA}, maHluwl_{TUN})$ '*opened*', both words are correct.
2. exaggerate form : (فتّاح, حلّال) $(fattAH_{MSA}, HallAl_{TUN})$, in this case TUN noun is wrong but the MSA noun is correct '*conqueror*'.
3. noun of place : (مفتح, محلّ) $(maftaH_{MSA}, mHall_{TUN})$, in this case TUN noun is correct '*shop, store*' while the MSA noun does not exist. The TUN noun is obtained after the application of the gemination[3] rule. The allows deleting the vowel between the second and the third radical.
4. analogous adjective : (فتيح, محلال) $(ftiyH_{MSA}, miHlAl_{TUN})$, both nouns are wrong.

Situations (3) and (4) can be handeled by filtering the MSA part using an MSA resource. In order to do so, we have used three resources :

- an Arabic corpus made of reports of the French Press Agency (AFP), which contains 1.5 million word forms. From these words, we have extracted $10,595$ types of nominal lemmas using the Arabic morphological analyser MADA (Habash et al., 2009). Only pairs that have the MSA noun in the corpus have been kept. At the end of this stage, we have obtained a lexicon of 20130 entries : 8441 MSA nouns and 2636 TUN nouns.
- an MSA large-scale lexicon SAMA (Graff et al., 2009) containing $36,935$ nominal lemmas. Our resulting lexicon contains $26,486$ entries : $4,712$ TUN nouns and $10,647$ MSA nouns.
- The union of these resources containing $40,172$ nominal lemmas. Using this resource, a lexicon made of $39,793$ was obtained : $5,017$ TUN nouns and $14,804$ MSA nouns. All results are given in section 4.

4 Evaluation

In order to evaluate the resource produced, we used a Tunisian corpus made of 800 sentences. In order to cover most spoken TUN varieties, the data was obtained from several sources: TV series, political debates, and a transcribed theater play (Dhouib, 2007). Once manually tokenized and annotated with morphological information (lemma and part-of-speech tag), the corpus contains $6,123$ tokens: 53.8% $(3,295)$ of them are nouns, among which 52% are deverbals.

We have divided the evaluation corpus into two different sets : a development corpus contaning 300 TUN sentences and a test corpus with 500 sentences.

Two metrics have been used to evaluate the deverbal lexicon produced. The first one is coverage, which is the part of the deverbal types of the evaluation corpus that are present in the lexicon. The second one is ambiguity which is the average number of target deverbals for a source deverbal.

There are two sources of ambiguity:

[3]The second and the third root radical are identical.

- The verbal lexicon can associate for one input verb many target verbs, for example the TUN verb مشى *mšý* matches with two different MSA verbs مشى *mšý 'to walk'* and ذهب *ðhb 'to go'*. The ambiguity is more important in the TUN → MSA sense. On average, a TUN pair corresponds to 1.78 MSA pairs, 1.11 in the opposite direction. The maximum ambiguity is equal to four in the MSA → TUN direction and sixteen in the opposite direction.
- the TUN-MSA deverbal table may define several patterns for a deverbal noun as shown in table 2.

The evaluation[4] of the deverbal lexicon on the test set is displayed in Table 3. The table shows that, without filtering the lexicon coverage is equal to 67.23%. Ambiguity (in the TUN→MSA direction) is equal to 12.58, which means that, on average, for a TUN deverbal, 12.58 MSA deverbals are produced. After filtering using AFP corpus, coverage drops to 60.04% and ambiguity to 6.99. Filtering with the SAMA lexicon yields a coverage of 62.66% and an ambiguity of 7.24. Finally, filtering using AFP ∪ SAMA, the coverage reaches 65.67% and the whith an ambiguity of 7.35.

filtering method	number of entries	coverage	ambiguity rate	
			MSA→TUN	TUN→MSA
none	173,407	67.23	7.65	12.58
AFP	17,896	60.04	2.36	6.99
SAMA	33,271	63.89	3.45	7.24
AFP ∪ SAMA	35,792	65.67	2.59	7.35

Table 3: Results on test set

As in the verbal lexicon, switching from TUN to MSA is more ambigous than the inverse direction. Ambiguity rates attests that MSA is lexically richer than TUN. The filtering step helps to significantly decrease ambiguity, but it also decreases coverage! The best result is the union of AFP∪SAMA, which enables us to obtain the best trade-off.

Table 4 summarizes the coverage and the ambiguity rate of the deverbal lexicon in the development and the test sets respectively :

filtering method	number of entries	coverage	ambiguity rate	
			MSA→TUN	TUN→MSA
none	173,407	66.12	7.65	12.58
AFP	17,896	59.23	2.36	6.99
SAMA	33,271	62.66	3.45	7.24
AFP ∪ SAMA	35,79	64.59	2.59	7.35

Table 4: Results in the development set

We have carried out an error analysis on the automatically generated lexical entries. There are three major causes that can explain a missing target deverbal:

1. Absence of the corresponding verb in the verbal lexicon: nouns deriving from a verb that is absent from the verb lexicon are not produced in the deverbal lexicon.
2. Missing entries in the TUN-MSA deverbal table
3. Missing morphological and orthographic rules.

In order to estimate the part of missing deverbals that is due to lack of coverage of the verbal lexicon, we have added verbs that derive missing deverbals of the development corpus. 92 verbal entries have been added. Table 5 shows results of coverage and ambiguity on the development set. This result, although artificial allows to compute an upper bound that can be attained with a more complete verbal lexicon.

As one can see in Table 5, coverage jumps from 66.12% to 87.33% before filtering and from 64.59% to 84.16% after filtering using AFP ∪ SAMA. The ambiguity rate increases slightly.

[4]In this paper, we don't use precision and recall measures because of the small size of the reference corpus.

filtering method	number of entries	coverage	ambiguity rate	
			MSA→TUN	TUN→MSA
none	195,917	87.33	7.93	12.86
AFP	20,130	81.46	2.24	7.17
SAMA	36,935	82.97	3.67	8.03
AFP ∪ SAMA	39,763	84.16	2.86	8.15

Table 5: Results in the development set after enriching the verbal lexicon

Table 6 gives the results obtained on the test set after enriching the verbal lexicon using the development set.

filtering method	number of entries	coverage	ambiguity rate	
			MSA→TUN	TUN→MSA
none	195,917	72.95	7.93	12.86
AFP	20,130	65.86	2.24	7.17
SAMA	36,935	68.41	3.67	8.03
AFP ∪ SAMA	39, 763	71.18	2.86	8.15

Table 6: Results in the test set after enriching the verbal lexicon

As shown in table 6, enriching the verbal lexicon improves significantly the coverage of the deverbal lexicon on the test set. In fact, it rises from 67% to 73% before filtering and from 65% to 71% after filtering using AFP∪SAMA, whereas ambiguity remains stable.

5 Root lexicon and pattern correspondance table

The previous section shows that a large portion of errors came from the lack of coverage of the verbal lexicon. By adding 92 verbal entries, the coverage jumps by about 6%. Among these 92 entries, there were 28 inexistent roots but for the 64 remaining, the root was already present in the verbal lexicon, we have just added new patterns to the roots (as the pair did not exist).

Sebsequently, we have divided the verbal lexicon into two independant resources : a root lexicon and a verbal pattern correspondance table.

The root lexicon is made of pairs of the form (r_{MSA}, r_{TUN}), where r_{MSA} is an MSA root and r_{TUN} is a TUN root. The root lexicon contains 1,357 entries. The MSA side contains 1,068 distinct roots and the TUN side 665 ones. 523 entries are composed of the same root on both sides. As in the verbal lexicon, the ambiguity is higher in the TUN → MSA direction. On average, a TUN root is paired with 2.07 MSA roots. In the opposite direction, 1.27 roots.

The verbal pattern correspondance table indicates, for a pattern in MSA or TUN, the most frequent corresponding pattern on the other side.

In this approach, the target pattern is selected by a lookup in the verbal pattern correspondance table but the target roots are selected by a root lexicon lookup. For each source root, we have combined it with all the nominal patterns corresponding to each verbal pattern. The target deverbal is made of the target root given by the lexicon root and the target nominal pattern depends on the target verbal pattern indicated in the verbal pattern correspondance table as shown in figure 2.

Results of this experiment on the test corpus show that using this method increase greatly the coverage. Although it also raises the number of generated entries and subsequently ambiguity.

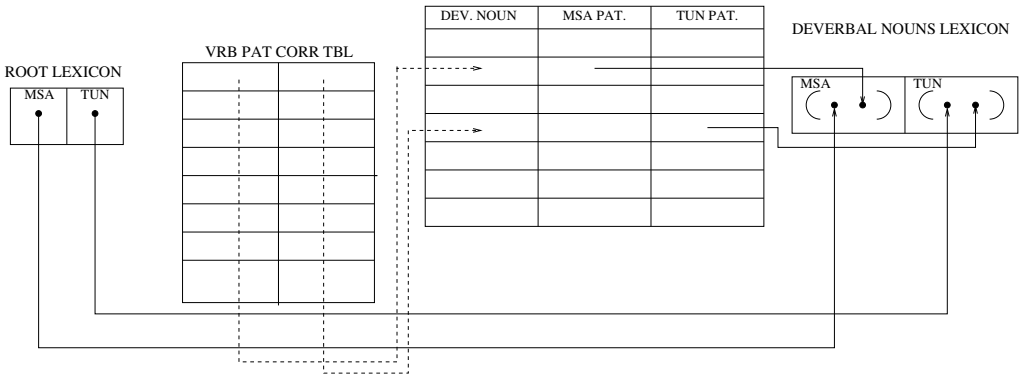

Figure 2: Generating TUN-MSA pairs of deverbal nouns using roots

filtering method	number of entries	coverage	ambiguity rate	
			MSA→TUN	TUN→MSA
no filtering	1,324,073	79, 13	18.47	36.42
filtering by AFP	122,315	71.33	6.66	31.04
filtering by SAMA	225,835	74.86	10.33	28.35
filtering by AFP ∪ SAMA	242, 104	76.83	6.57	28.68

Table 7: TUN-MSA Deverbal Table

6 Conclusion and Future Work

In this paper, we have presented a bilingual lexicon of deverbal nouns between MSA and TUN. Our method aims to extend an existing TUN verbal lexicon using a table of deverbal patterns to automatically generate pairs of TUN and MSA deverbal nouns. Several MSA resources were used to filter wrong pairs generated. The lexicon was evaluated using two metrics: coverage and ambiguity.

The coverage given by our lexicon is about 71%. Ambiguity is slightly high in TUN→MSA direction. It reaches 8.15. A contextual disambiguation process is therefore necessary for such a process to be of practical use.

In future work, we plan to include this lexicon into a system of translation from TUN to an approximative form of MSA which will be parsed using an MSA parser.

References

Mustafa Al-Ghulayaini. 2010. جامع الدروس العربية *jAmς Aldrws Alςrbyħ, Part II*. IslamKotob.

Rahma Boujelbane, Meriem Ellouze Khemekhem, and Lamia Hadrich Belguith. 2013. Mapping rules for building a tunisian dialect lexicon and generating corpora.

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

Elmoncef Dhouib. 2007. *El Makki w-Zakiyya*. Publishing House Manshuwrat Manara, Tunis, Tunisia.

C.A. Ferguson. 1959. Diglossia. *Word*, 15(2).

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

N. Habash and O. Rambow. 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.

N. Habash, O. Rambow, and G. Kiraz. 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.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

N. Habash, R. Eskander, and A. Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In *In proceedings of Traitement Automatique du Langage Naturel (TALN 2013)*.

Mohamed Maamouri, Tim Buckwalter, and Christopher Cieri. 2004. Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools*.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of NAACL-HLT*, pages 348–358.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.