# *Dialectal Machine Translation*
# *A Case Study*

**LABED Zohra**
University of Mostaganem

**Résumé**
Traditional grammarians are well-known for their clear standing position against the spoken form of language. Primacy was, at their time, orientated towards writing which was believed to represent, unlike speech, the pure and correct version of language. Their attitude was previously greatly influential. It was until the advent of structural linguistics that language scholars started to claim the significance of studying the spoken form. Speech was recognised as an autonomous object of investigation, but not its variation. First linguists devoted their attention to stable linguistic features and neglected those which vary. Variability meant for them disorder and chaos. Language variation has however started to gain meditation since sociolinguistics appeared. That language variation is systematic and worth exploring has been realised gradually. Today, sociolinguists hold against any kind of variable form marginalisation. Needless to say that the most salient language variation is regional and lexical. One objective of sociolinguists is to promote the spoken varieties or dialects, and provide them with the necessary tools already supplied to the written form of language. Cooperation between language scholars and computer scientists can turn automatic translation possible to any dialectal variety. Given that dialects are particularly exposed to lexical variation, the question which arises here: into which item is the computer supposed to translate? Our choice has fallen on Oran Arabic. In this paper, we will attempt to find an answer to this question on fieldwork bases.

## 1. Introduction

Population movements (or internal migration) have always occurred throughout human life. The industrial revolution has played the role of accelerating the process from rural to urban areas in the western countries. The same case has obtained in the third world countries, and more remarkably after their independence. Migration within the post-independent Algeria has led to contact between the newcomers' varieties[1] (mainly Temouchent (TMT); Mascara (MKR); Sidi Bel Abbes (SBA); Tiaret (TRT); Bayad (BYD); Biskra (BSK); Bechar (BSR); Mostaganem (MST); Saida (SAD); Nedroma (NDM); Tlemcen (TSN) (see Bouamrane 1991, 1993) and between the latter and the local variety. Additional linguistic forms have been introduced to urban Algerian Arabic. This has induced enormous linguistic variation within urban Arabic systems. However, the notion 'dialect' usually refers to the kind of Arabic employed in daily conversations, whereas the notion 'language' indicates Classical Arabic[2] (henceforth SA) which is officially recognised by the national constitution. Broadly speaking, sociolinguists have often pronounced crucial criticisms of the way distinction between languages and dialects is made. In contrast with a language, a dialect is frequently viewed as a non-standard unwritten variety. Due to extra-linguistic factors, one kind of language is attributed the designation 'language' or 'dialect. Linguistically however, all the varieties are equal. Standard written language forms are usually used for translation. If we take the sociolinguistic viewpoint into account, will it be possible to translate into or from a non-standard unwritten variety? Is translation workable in the presence of linguistic (particularly lexical) variation?

## 2. The Linguistic Impacts of Globalisation

Economically, globalisation means "… the way in which processes of production and consumption, and the consequent flows of capital, operate increasingly on a global, rather than 'local' or national levels." (Swann et al, 2004: 127). The process, in other words, refers to the internationalisation of

[1] "… a neutral term to apply to any 'kind of language' we wish to talk about without being specific" (Trudgill, 2000: 05)

[2] or one of its versions such as Modern Standard Arabic

manufacturing, trade and benefits conducted by transnational partnerships. As a member of the Organisation of the Petroleum Exporting Countries (OPEC), Algeria is seen as one of the African countries that major in the hydrocarbon sector. Oil and natural gas production has been key to its economic growth. In the last decades, the country fostered the expansion of foreign participation in this sector exploration. In response to the increasing global system challenges, Algeria has recently undertaken further important economic reforms to adapt and find a place in the contemporary world. Apart from the energy sector, successive Algerian governments have found it paramount to broaden the country’s international corporations to include other various economic fields. Foreign investments have received official welcomes to proceed in industry, housing (such as flat buildings) and road constructions (such as motorways and tramway projects). Indeed, Algeria has been the theatre of receiving huge remarkable numbers of foreigners such as the Chinese, Turkish, Spanish, French, British, Americans and many others as long-term investors. This econo-demographic situation has voiced reconsiderations of the new communicative needs. The question which arises: how can these foreigners daily interact with the Algerians? We do not here refer to those bilinguals from either parts, but many foreign and local labours (who have not undergone a considerable education) who come in contact, and speak their mother tongues solely. Or, it happens that a monolingual Chinese or Turkish labour is engaged in a face-to-face interlocution with an Algerian seller who does not master their varieties. Does SA learning fulfill their ordinary requirements? It would be very useful if this learning takes place at the writing level. But regarding speaking, SA is the mother tongue of nobody. The Algerian majority uses Dialectal Arabic in everyday life. Those foreigners usually and apparently need this kind of language to meet their communicative necessities. We do not exclude either the Arab investors who come from the

Middle East; many of them complain of the fact they find difficulties in understanding Algerian dialects. Their interactions with the Algerians are, for them, highly characterised by unintelligibility. Their SA use is possible, but still very artificial and does not fulfill their everyday requirements. Once again, they need to get familiar with the local Dialectal Arabic. In this paper, we support the tendency orientated towards dialectal translation. Indeed, translating from one or into a dialect may greatly facilitate daily communication, and realise therefore effective partnerships. This type of work can help in collaborating with computer scientists to find suitable ways for automatic translation from and/or into Algerian dialects. This type of translation, on the other hand, requires thorough understanding of the dialectal linguistic composition via scientific and well-defined studies. We currently take the illustrative case of Oran dialectal lexis. Sociolinguistic observations show that the dialect in question is subjected to lexical variation: one concept may have different lexical variants. Our main research question is: Which items will be provided to computer translation? In other words, what are the most frequent lexical variants in Oran dialect (henceforth ORD)? Due to space-limitations, we will be content with seven basic verbal forms. An oral directive anonymous questionnaire (a set of French items to be translated into ORD) was then supplied to 79 university informants who were all Oranees[3]. They were thirty seven males and forty two females aged between nineteen and twenty seven years old. We are particularly interested in the extra-linguistic variable ‘region’ which meets the current research question.

## 3. Our Findings

### 3.1. BE ABLE TO

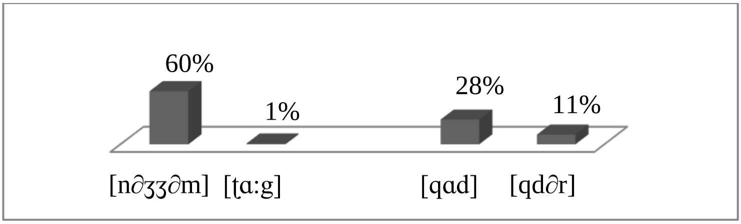
[3] They were born in Oran

Four lexical variants were produced by the informants once asked to translate the French form /puvwaR/ *be able to* into ORD. The results attained are represented as follows,

| **utterance** | **frequency** | **percentage** | **gloss** |
|---|---|---|---|
| [n∂ʒʒ∂m] | 47 | 60% | *be able to* |
| [ţɑ:g] | 01 | 01% | |
| [qɑd] | 22 | 28% | |
| [qd∂r] | 09 | 11% | |

Table 1: ORD Variants with the meaning be able to

[n∂ʒʒ∂m]-realisation (60%) overtakes that of [qɑd] (28%). The percentages of [qd∂r] and [ţɑ:g]-articulators, (11%) and (01%) respectively, record inferior results. It is seemingly the variant [n∂ʒʒ∂m] which ranks first, followed by [qɑd] which originally featured ORD[4]. The former outweighes the other three items probably because [n∂ʒʒ∂m] obtains in the migratory dialectal majority (MST, MKR, SBA, SAD, BSR) (Bouamrane, 1993). Competition between the lexical variants [n∂ʒʒ∂m] and [qɑd] seemingly persists within the dialectal mix. We predict [n∂ʒʒ∂m] to finally vanquish in case no serious extra-linguistic changes (such as further demographic mobility) takes place in Oran. The bare-graph below corresponds to the table above,



[4] It also originated from TMT (Bouamrane, 1993), possibly because TMT was part of ORD in the near past.

Figure 1: ORD Variants with the meaning be able to

### 3.2. GET DOWN

As part of the oral questionnaire, another French form, /desãdR/ *get down*, was presented for dialectal translation. The following table demonstrates that [hæw∂d] scores in the first position (75%) whereas the second rank (24%) is realised by [nz∂l]. [hbɑʈ] records only the percentage of 01%. Those Oranees relatively avoid using dialectally [nz∂l] and [hbɑʈ] for stereotypical reasons. [nz∂l] corresponds to CA /nazala/ often heard in formal political or educational environments while [hbɑʈ][5] is widely believed to reflect the Algerian capital's speech which is quiet distinct from ORD. Relatively, [hæw∂d] is less stereotypically salient and normalised in the local dialect. Our table is followed by an illustrative figure,

| **utterance** | **frequency** | **percentage** | **gloss** |
|---|---|---|---|
| [nz∂l] | 19 | 24% | *get down* |
| [hæw∂d] | 59 | 75% | |
| [hbɑʈ] | 01 | 01% | |

Table 2: ORD Variants with the meaning *get down*

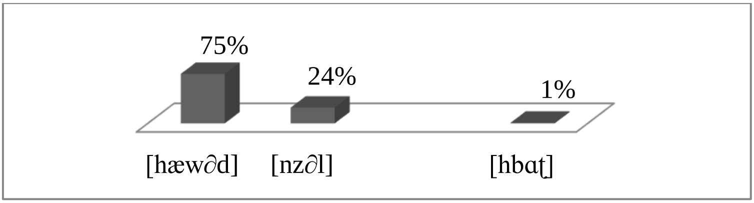


Figure 2: ORD Variants with the meaning get down

### 3.3. GO

The informants provided four lexical variants when translating the French term /paRtiR/. [ra:ħ]-production very clearly

[5] More exactly [ħbɑʈ]

preponderates (95%). The remaining 5% is shared by the other three produced forms, namely [rawwæħ] (03%), [ʃ∂ww∂r] (01%), [mʃa] (01%). The following table displays that most of the lexical variants are, unlike the predominant stabilised [ra:ħ], in their way of disappearing.

| utterance | frequency | percentage | gloss |
|---|---|---|---|
| [ra:ħ] | 75 | 95% | *go* |
| [rawwæħ] | 02 | 03% | |
| [ʃawwar] | 01 | 01% | |
| [mʃa] | 01 | 01% | |

Table 3: ORD Variants with the meaning *go*

Indeed, the local variety involves [ra:ħ], together with the transplanted dialects TRT, MKR, MST, SAD (see Bouamrane, 1993). On the other hand, [ʃ∂ww∂r] characterises BSA; [rawwæħ] reflects TMT; and [mʃæ] is encountered in TSN and NDM (ibid). What happens, in fact, is not a variantal eradication, but linguistic functional reattribution: the variants [ʃ∂ww∂r] and [mʃa] are rather re-assigned other different semantic roles. [ʃ∂ww∂r] fulfills the meaning of *leave* while [mʃa] means *walk*. [rawwæħ] is exceptionally simplified to [ra:ħ]. The results are exemplified in the graph below,

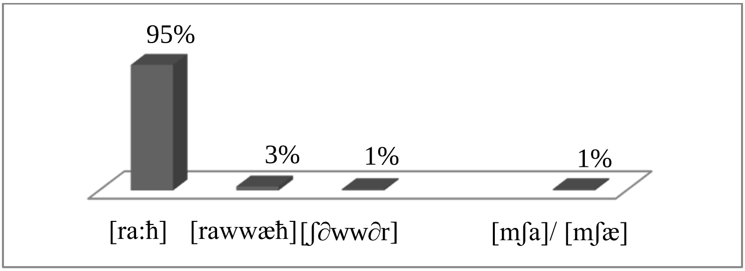


Figure 3: ORD Variants with the meaning go

3.4. HIDE

Our informants responded to our requirement of translating the French item /kaʃe/ *hide*, and realised four lexical variants grouped in the table below,

| utterance | frequency | percentage | gloss |
|---|---|---|---|
| [χz∂n] | 61 | 78% | *hide* |
| [(n)d∂s] | 07 | 09% | |
| [dr∂g] | 02 | 02% | |
| [t(s)χ∂bba] | 09 | 11% | |

Table 4: ORD Variants with the meaning hide

The following figure gives a clearer picture of the results arrived at,

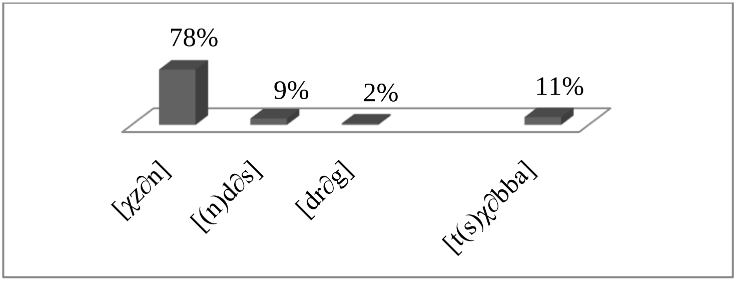


Figure 4: ORD Variants with the meaning hide

The preponderant item ([χz∂n]) is again related to the fact it prevails in the speech of the greatest deal of users. According to Bouamrane (1993), TRT, TMT, MKR and MST cover the variant which also originally takes place in ORD. Although the other lexical elements also mark their regional affiliation: ([(n)d∂s] is found in BYD, BSR, SBA; SAD includes [dr∂g]; and [t(s)χ∂bba] is comprised in TSN and NDM), they are only minor in number in our findings, probably due to the inferior number of their migratory users.

3.5. HOLD

The translation of the French conjugated verbal forms /il tjɛ̃/ *he holds* engendered a number of lexical variants tabled below.

| utterance | frequency | percentage | gloss |
|---|---|---|---|
| [ʃ∂d] | 43 | 54% | *hold* |
| [ħk∂m] | 10 | 13% | |
| [gbɑɖ] | 07 | 9% | |
| [gɖɑb] | 19 | 24% | |

Table 5: ORD Variants with the meaning *hold*

The greatest percentage (54%) belongs to [ʃ∂d]. The set of variants [ħk∂m], [gbɑɖ] and [gɖɑb] rank in lower positions. This lexical situation emerges from the fact that the previously native [ħk∂m] is being diminished (13%) under the impact of [ʃ∂d], a variant transferred to the Oranees through the enormous population mobility towards their town. However, [gbɑɖ] is still being metathesised, and possibly disappearing. [gɖɑb] is in contrast arising, but with a distinct semantic role: It is more and more employed as *catch*. Therefore, two items are conserved in the sense that [ʃ∂d] conveys the meaning of *hold* while [gɖɑb] opts for the meaning *catch*. Again from the table, we notice divergent scores of the two variants. We predict their convergence over time.

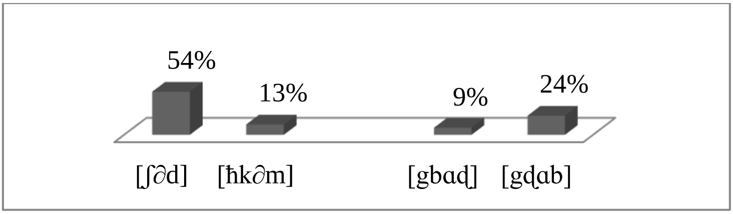


Figure 5: ORD Variants with the meaning hold

3.6. LIE

Consider the following table. A great deal of lexical variants was elicited from the respondents, a fact that illustrates a typical mix linguistic situation in Oran,

| utterance | frequency | percentage | gloss |
|---|---|---|---|
| [tk∂ss∂l] | 55 | 70% | *lie down* |
| [tm∂d(∂d)] | 02 | 02% | |
| [(t)war∂k] | 03 | 04% | |
| [r∂jjæħ] | 05 | 07% | |
| [ṭ(ɑ:)lɑg (ruħæh)] | 08 | 10% | |
| [tw∂ka] | 03 | 04% | |
| [mɑṣṭɑle] | 01 | 01% | |
| [ṭɑrɑħ] | 01 | 01% | |
| [qɑjɑṣ ruħæh] | 01 | 01% | |

Table 6: ORD Variants with the meaning lie down

Bouamrane (1993) indicates [tk∂ss∂l] as originally local item encountered concurrently in TRT and MKR. [tm∂d] comes from BSR, BYD, SBA (one participant provided the variant as [tm∂d∂d]). The item [(t)war∂k] originates from TSN and NDM. The remainder minor number of items produced respectively as (02%), (07%), (10%), (04%), (01%), (01%), (01) has no clear regional affiliation. Their corresponding [tm∂d(∂d)], [r∂jjæħ], [ṭ(ɑ:)lɑg ruħæh], [tw∂ka], [mɑṣṭɑle], [ṭɑrɑħ], [qɑjɑṣ ruħæh] are increasingly eradicated utterances. Despite this complex linguistic situation, [tk∂ss∂l] still preponderates (70%). The following graph epitomises the above table,

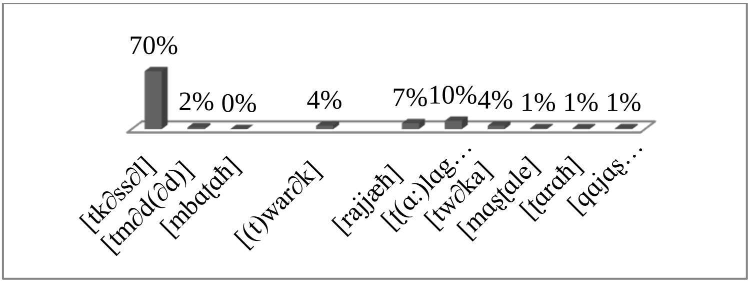


Figure 6: ORD Variants with the meaning lie down

## 3.7. SEND

Our results reflect the participants' translation of the French item /ãvwaje/ *send*. Examining the following table and its bare-graph reveals that [rs∂l] is uttered by almost the three quarters of the whole number of informants. Yet, the utterances [ze:fɑʈ],[ʂe:fɑʈ] and [bʕæt] record inferior percentages (not more than 4%) in terms of frequency.

| utterance | frequency | percentage | gloss |
|---|---|---|---|
| [rs∂l] | 74 | 94% | *send* |
| [ze:fɑʈ] | 01 | 01% | |
| [ʂe:fɑʈ] | 01 | 01% | |
| [bʕæt] | 03 | 04% | |

Table 7: ORD Variants with the meaning *send*

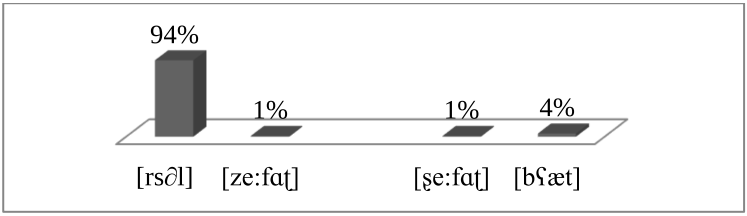


Figure 7: ORD Variants with the meaning send

Basically, ORD (together with TMT and SBA) previously embraced the feature [ze:fɑʈ] (Bouamrane, 1993). This variant is articulated with an initial alternative voiceless sound ([ʂe:fɑʈ]) in other dialects (such as TSN). The two items, nevertheless, are almost vanishing in our informants' speech production. The ground is rather left to the highest percentage scored by [rs∂l] (94%). This finding refers to an item which originates from TRT, BSR, BYD, MKR (ibid). Both [rs∂l] and [bʕæt] correspond respectively to CA /ʔarsala/ and /baʕaθa/. Yet, the first utterance seems more dialectal since it is used by the largest majority of migrants, in contrast to the latter which appears more formal.

## Conclusion

In this paper, we have tried to turn attention to those language forms which are nearly ignored by machine translation. Standard written varieties have constituted the meeting point of collaboration between computer scientists and linguists. Their efforts have been certainly but not totally fruitful. Mother tongues, which are non-standard unwritten forms, are still very important in fulfilling everyday requirements which is not always the case of standard written varieties. Yet, their automatic translation emanates only a little. Such a situation is encountered in the Arab World. Algeria is a good example for this investigation. Particularly, Oran receives large numbers of foreign workers who need to meet their daily communicative requirements. Dialectal machine translation could be very useful in detecting the appropriate utterance for the relevant situation. Since ORD is exposed to language variation, our oral questionnaire has helped to display the most frequent items in the local dialect. Due to space-restrictions, we have not, in this paper, gone beyond seven verbal variants which are [n∂ʒʒ∂m] *be able to,*[hæw∂d] *get down*, [ra:ħ] *go*, [χz∂n] *hide*, [ʃ∂d] *hold*, [tk∂ss∂l] *lie down*, [rs∂l] *send*. The lexical variantal list is still long. We have some complementary data at hand that we tend to further explore in a future paper.

## *References*

Bouamrane, A (1991) Lexical variation among Arabic Dialects in Algeria. *Cahiers de Dialectologie et de Linguistique Contrastive*, vol.2, ILE-Oran, Algeria.

Bouamrane, A (1993) ‘More on Lexical Variation among Arabic Dialects in Algeria’. In *Cahiers de Dialectologie et de Linguistique Contrastive*. Vol, IV. N°1. 15-27.

Swann et al (2004) *A Dictionary of Sociolinguistics*. Tuscaloosa: The University of Alabama Press

Trudgill, P (2000) *Sociolinguistics: an Introduction to Language and Society*. ed. U.K: Longman