An Algerian dialect: Study and Resources
Salima Harrat, Karima Meftouh, Mourad Abbas, Walid-Khaled Hidouci, Kamel Smaïli

To cite this version:
Salima Harrat, Karima Meftouh, Mourad Abbas, Walid-Khaled Hidouci, Kamel Smaïli. An Algerian dialect: Study and Resources. International journal of advanced computer science and applications (IJACSA), 2016, 7 (3), pp.384-396. 10.14569/IJACSA.2016.070353. hal-01297415

HAL Id: hal-01297415
https://hal.science/hal-01297415v1
Submitted on 5 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
An Algerian dialect: Study and Resources

Salima Harrat*, Karima Meftouh†, Mourad Abbas‡, Khaled-Walid Hidouci§ and Kamel Smaili¶
*Ecole Supérieure d’Informatique (ESI), Algiers, Algeria
†Badji Mokhtar University, Annaba, Algeria
‡CRSTDLA Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe, Algiers, Algeria
§Ecole Supérieure d’Informatique (ESI), Algiers, Algeria
¶Campus Scientifique LORIA, Nancy, France

Abstract—Arabic is the official language overall Arab countries, it is used for official speech, newspapers, public administration and school. In Parallel, for everyday communication, non-official talks, songs and movies, Arab people use their dialects which are inspired from Standard Arabic and differ from one Arabic country to another. These linguistic phenomena is called diglossia, a situation in which two distinct varieties of a language are spoken within the same speech community. It is observed throughout all Arab countries, standard Arabic widely written but not used in everyday conversation, dialect widely spoken in everyday life but almost never written. Thus, in NLP area, a lot of works have been dedicated for written Arabic. In contrast, Arabic dialects at a near time were not studied enough. Interest for them is recent. First work for these dialects began in the last decade for middle-east ones. Dialects of the Maghreb are just beginning to be studied. Compared to written Arabic, dialects are under-resourced languages which suffer from lack of NLP resources despite their large use. We deal in this paper with Arabic Algerian dialect a non-resourced language for which no known resource is available to date. We present a first linguistic study introducing its most important features and we describe the resources that we created from scratch for this dialect.

Keywords—Arabic dialect, Algerian dialect, Modern Standard Arabic, Grapheme to Phoneme Conversion, Morphological Analysis

I. INTRODUCTION

Under-resourced languages are languages which lacks resources dedicated for natural language processing. In fact, these languages suffer from unavailability of basic tools like corpora, mono or multilingual dictionaries, morphological and syntactic analyzers, etc. This lack of resources makes working with these languages a great challenge, especially when we deal with unwritten languages like Arabic dialects. Compared to other under-resourced languages, Arabic dialects present the following additional difficulties:

• Since they are spoken languages they are not written and there are no established rules to write them. A same word could have many orthographic forms which are all acceptable since there is no writing rules as reference.

• The flexibility in the grammatical and lexical levels despite their belonging to Arabic Language.

• Besides the fact that these dialects are different from Arabic, they are also different from each other. For instance, dialects of the Maghreb differ from those of the middle-east. They may be also different inside the same country.

• These dialects are also widely influenced by other languages such as French, English, Spanish, Turkish and Berber.

In Algeria, as well as in all arab countries, these dialects are used in everyday conversations. However, with the advent of the internet they are increasingly used in social networks and forums. They emerge on the web as a real communication language due to the ease to communicate in dialect especially for people with low level of education. But unfortunately basic NLP tools for these dialects are not available.

This work is a first part of the Project TORJMAN which is a Speech-To-Speech Translator between Algerian Arabic dialects and MSA. Unlike Middle-East Arabic dialects, Algerian Arabic dialects are non-resourced languages, they lack all kinds of NLP resources. Consequently, TORJMAN begins from Scratch.

In this paper, we describe and extend resources creation tasks for Arabic dialect of Algeria that appeared in [1] and [2]. We focus on Algiers dialect which is the spoken Arabic of Algiers (capital city of Algeria) and its periphery. This choice is justified by the fact that this dialect is the one we know best and practice since we are native speakers of this dialect. For convenience of reference, we will design Algiers dialect by ALG, this will make this manuscript easier to read.

This paper is organized as follows: before dealing with Algerian dialect we give in Section II a brief overview of Arabic language, whereas in Section III we present different aspects of ALG. The following Sections will be dedicated to the resources that we created, we detail how we made the first corpus of Algiers dialect (Section IV). Then we present ALG grapheme-phoneme converter (Section V) which has allowed us to get a phonetized corpus of Algiers dialect. In Section VI, we describe how we created a morphological analyser for ALG by adapting BAMA the well known analyser for MSA. Finally, we will conclude by summarizing the main ideas of this work and by giving our future tendencies.

II. ARABIC LANGUAGE

Arabic is a Semitic language, it is used by around 420 million people. It is the official language of about 22 countries. Arabic is a generic term covering Arabic is the official language overall Arab countries, it is used for official speech, newspapers, public administration and school. In Parallel, for everyday communication, non-official talks, songs and movies, Arab people use their dialects which are inspired from Standard Arabic and differ from one Arabic country to another. These linguistic phenomena is called diglossia, a situation in which two distinct varieties of a language are spoken within the same speech community. It is observed throughout all Arab countries, standard Arabic widely written but not used in everyday conversation, dialect widely spoken in everyday life but almost never written. Thus, in NLP area, a lot of works have been dedicated for written Arabic. In contrast, Arabic dialects at a near time were not studied enough. Interest for them is recent. First work for these dialects began in the last decade for middle-east ones. Dialects of the Maghreb are just beginning to be studied. Compared to written Arabic, dialects are under-resourced languages which suffer from lack of NLP resources despite their large use. We deal in this paper with Arabic Algerian dialect a non-resourced language for which no known resource is available to date. We present a first linguistic study introducing its most important features and we describe the resources that we created from scratch for this dialect.

Keywords—Arabic dialect, Algerian dialect, Modern Standard Arabic, Grapheme to Phoneme Conversion, Morphological Analysis

I. INTRODUCTION

Under-resourced languages are languages which lacks resources dedicated for natural language processing. In fact, these languages suffer from unavailability of basic tools like corpora, mono or multilingual dictionaries, morphological and syntactic analyzers, etc. This lack of resources makes working with these languages a great challenge, especially when we deal with unwritten languages like Arabic dialects. Compared to other under-resourced languages, Arabic dialects present the following additional difficulties:

• Since they are spoken languages they are not written and there are no established rules to write them. A same word could have many orthographic forms which are all acceptable since there is no writing rules as reference.

• The flexibility in the grammatical and lexical levels despite their belonging to Arabic Language.

• Besides the fact that these dialects are different from Arabic, they are also different from each other. For instance, dialects of the Maghreb differ from those of the middle-east. They may be also different inside the same country.

• These dialects are also widely influenced by other languages such as French, English, Spanish, Turkish and Berber.

In Algeria, as well as in all arab countries, these dialects are used in everyday conversations. However, with the advent of the internet they are increasingly used in social networks and forums. They emerge on the web as a real communication language due to the ease to communicate in dialect especially for people with low level of education. But unfortunately basic NLP tools for these dialects are not available.

This work is a first part of the Project TORJMAN which is a Speech-To-Speech Translator between Algerian Arabic dialects and MSA. Unlike Middle-East Arabic dialects, Algerian Arabic dialects are non-resourced languages, they lack all kinds of NLP resources. Consequently, TORJMAN begins from Scratch.

In this paper, we describe and extend resources creation tasks for Arabic dialect of Algeria that appeared in [1] and [2]. We focus on Algiers dialect which is the spoken Arabic of Algiers (capital city of Algeria) and its periphery. This choice is justified by the fact that this dialect is the one we know best and practice since we are native speakers of this dialect. For convenience of reference, we will design Algiers dialect by ALG, this will make this manuscript easier to read.

This paper is organized as follows: before dealing with Algerian dialect we give in Section II a brief overview of Arabic language, whereas in Section III we present different aspects of ALG. The following Sections will be dedicated to the resources that we created, we detail how we made the first corpus of Algiers dialect (Section IV). Then we present ALG grapheme-phoneme converter (Section V) which has allowed us to get a phonetized corpus of Algiers dialect. In Section VI, we describe how we created a morphological analyser for ALG by adapting BAMA the well known analyser for MSA. Finally, we will conclude by summarizing the main ideas of this work and by giving our future tendencies.

II. ARABIC LANGUAGE

Arabic is a Semitic language, it is used by around 420 million people. It is the official language of about 22 countries. Arabic is a generic term covering all kinds of NLP resources. Consequently, TORJMAN begins from Scratch.

In this paper, we describe and extend resources creation tasks for Arabic dialect of Algeria that appeared in [1] and [2]. We focus on Algiers dialect which is the spoken Arabic of Algiers (capital city of Algeria) and its periphery. This choice is justified by the fact that this dialect is the one we know best and practice since we are native speakers of this dialect. For convenience of reference, we will design Algiers dialect by ALG, this will make this manuscript easier to read.

This paper is organized as follows: before dealing with Algerian dialect we give in Section II a brief overview of Arabic language, whereas in Section III we present different aspects of ALG. The following Sections will be dedicated to the resources that we created, we detail how we made the first corpus of Algiers dialect (Section IV). Then we present ALG grapheme-phoneme converter (Section V) which has allowed us to get a phonetized corpus of Algiers dialect. In Section VI, we describe how we created a morphological analyser for ALG by adapting BAMA the well known analyser for MSA. Finally, we will conclude by summarizing the main ideas of this work and by giving our future tendencies.

II. ARABIC LANGUAGE

Arabic is a Semitic language, it is used by around 420 million people. It is the official language of about 22 countries. Arabic is a generic term covering 3 separate groups:

1$^{TORJMAN}$ is a national research project which is totally financed by the Algerian research ministry, this appellation means translator or interpreter in English.
385 | Page

III. SPECIFICITIES OF ALGIERS DIALECT

Algerian dialect (ALG) is the dialectical Arabic spoken in Algiers and its periphery. This dialect is different from the dialects spoken in the other places of Algeria. It is not used in schools, television or newspapers, which usually use standard Arabic or French, but is more likely, heard in songs if not just in Algerian homes and on the street. Algerian Arabic is spoken daily by the vast majority of Algerians [7]. ALG as the other Arabic dialects simplifies the morphological and syntactic rules of the written Arabic. In [8], the author draws how match spoken Arabic is different from written Arabic in various language levels: Phonological differences between Classical Arabic and spoken Arabic are moderate (compared to other pairs of language-dialect), whereas grammatical differences are the most striking ones. At lexical level, differences are marked with variations in form and with differences of use and meaning.

Indeed, at phonological level, ALG (naturally) shares the most features related to Arabic. In addition to the 28 consonants phonemes of Arabic (given in Table I), ALG consonantal system includes non Arabic phonemes like /l/ as in the word /ì(r)/ (all), and the phonemes /p/] and /w/ used mainly in words borrowed from French like the case of /ð/ (adapted from the French word ”bôme” which means a bag) and /T/ (adapted from the French word ”valise” which means a bag). Also, it should be noted that the use of the phonemes (ط) and (ص) is very rare, most of the time َط is pronounced /d/ (ص) and َص is pronounced /d/ (ص). The same case is observed for /ى/ (ى) which is pronounced /h/ (ى). Note that the last two substitutions are observed also for Jordanian dialect [9].

Phonological features of ALG will be detailed further in this paper (section V).

A. Vocabulary

Algerian dialect has a vocabulary inspired from Arabic but the original words have been altered phonologically, with significant Berber substrates, and many new words and loanwords borrowed from French, Turkish and Spanish. Even though most of this vocabulary is from MSA, there is significant variation in the vocalization in most cases, and the omission or modification of some letters in other cases (mainly the Hamza [9]). Vocabulary of Algiers’s dialect includes verbs, nouns, pronouns and particles. In the following a brief description of each category.

- Verbs

Some verbs in ALG can adopt entirely the same

| Letter | Phoneme | Letter | Phoneme | Letter | Phoneme |
|--------|---------|--------|---------|--------|---------|
| أ | /a/ | أ | /a/ | أ | /a/ |
| إ | /e/ | إ | /e/ | إ | /e/ |
| ي | /i/ | ي | /i/ | ي | /i/ |
| ء | /e/ | ء | /e/ | ء | /e/ |
| م | /m/ | م | /m/ | م | /m/ |
| ن | /n/ | ن | /n/ | ن | /n/ |
| ل | /l/ | ل | /l/ | ل | /l/ |
| ر | /r/ | ر | /r/ | ر | /r/ |
| د | /d/ | د | /d/ | د | /d/ |
| ذ | /z/ | ذ | /z/ | ذ | /z/ |
| ز | /z/ | ز | /z/ | ز | /z/ |

TABLE I: Arabic phonemes using SAMPA

2The influence of European languages is due to the fact that most of the Arab countries were European colonies during the 19th century.

3An other classification is given in [6] where rural and Bedouin Arabic dialects are distinguished because of ethnic and social diversity of Arabic speakers. The author states that Bedouin dialects tend to be more conservative and homogeneous, while urban dialects show more evolutive tendencies.
scheme of MSA verbs by respecting the same vocalization such as in the case of the verb \(\text{idā} \) (to name) or \(\text{idā} \) (to salute). Other verbs are pronounced differently from corresponding MSA verbs by adopting different diacritics marks as the case of the verbs \(\text{šu´r} \) (to drink). An other set of dialect verbs are obtained by the omission or modification of some letters. In Table II we give some examples of each listed case.

**TABLE II: Examples of verbs scheme differences between ALG and MSA.**

| ALG Verb | Corresponding MSA Verb | Meaning Situation | Situation |
|----------|------------------------|------------------|----------|
| علی | علی | To salute | Same scheme |
| قابل | قابل | To confront | Same diacritics marks |
| حبر | حبر | To drink | /Different Diacritics marks |
| کتیب | کتیب | To write | Letters omission or modification |
| کئل | کئل | To come | |
| کئل | کئل | To remain | |
| کئل | کئل | To eat | |
| کئل | کئل | To finish | |

Another set of ALG verbs are those borrowed from foreign languages especially French such as \(\text{شارجا} \) which corresponds to the French verb “charger” (to load) or قاّرًا a modification of the verb "garer" which means to park.

- **Nouns**
  - Arabic ALG nouns can be primitive (not derived from any verbal root) or derived from verbs like for verbal names and participles (active and passive), in Table III an exemple is given. We should note that ALG nouns include an important portion of french words. Most of them are the results of a wide phonological alteration of original words such as \"م‌وتور\" ("moteur" in French, motor ), \"ل‌اطل‌سین وون\" ("la tension", blood pressure) and \"پولسیز\" ("policiere", policeman). Nouns include also numbers which represent units, tens, hundreds, etc. From 1 to 10 the numbers are close to MSA (with different vocalization), except for the numbers 0 and 2: the first one is pronounced as in French /zero/, and the second is ل‌تین، whereas in MSA it is ل‌مین. From number 11 to 19 the pronunciation in ALG differs from MSA, some letters and diacritics change but the number can be perceived easily by an Arab speaker. Numbers greater than 20 are also close to MSA numbers, only the diacritics marks differ.

**TABLE III: Example of ALG nouns derived from a verb.**

| Verb | Verbal name | Active participle | Passive participle |
|------|-------------|------------------|-------------------|
| علی | علی | علی | علی |
| قابل | قابل | قابل | قابل |
| حبر | حبر | حبر | حبر |
| کتیب | کتیب | کتیب | کتیب |
| کئل | کئل | کئل | کئل |

include demonstrative and personal pronouns. For relative pronouns, there is only one in Algiers dialect which is ل‌ذهی (that); this pronoun is used for female, masculine, singular and plural. We give in Tables IV and V all ALG used pronouns. It is important to note that the dual in ALG does not exist; there are no equivalent for Arabic pronouns ل‌تینا (second person, dual) and ل‌هی (third person, dual). Similarly, personal pronouns relative to feminine plural ل‌تینان and ل‌هیان related to second and third person respectively do not exist.

**TABLE IV: Personal pronouns of Algiers dialect.**

| 1st Person | 2nd Person | 3rd Person |
|-----------|------------|------------|
| ل‌ذهی | ل‌ذهی | ل‌ذهی |
| ل‌ذهی | ل‌ذهی | ل‌ذهی |
| ل‌ذهی | ل‌ذهی | ل‌ذهی |

**TABLE V: Demonstrative pronouns of Algiers dialect.**

| Singular | Plural |
|---------|--------|
| ل‌ذهی | ل‌ذهی |
| ل‌ذهی | ل‌ذهی |
| ل‌ذهی | ل‌ذهی |
| ل‌ذهی | ل‌ذهی |

- **Particles**
  - Particles are used in order to situate facts or objects relatively to time and place. They include different categories such us: prepositions (في, in, علی, on, ِب, with), coordinating conjunctions (و, and, ام‌بعد, after), quantifiers (الي, all, ل‌مین, few ).

**B. Inflection**

Algiers dialect is an inflected language such as Arabic. Words in this language are modified to express different grammatical categories such as tense, voice, person, number, and gender. It is well-known that depending on word category, the inflection is called conjugation when it is related to a verb, and declension when it is related to nouns, adjectives or pronouns. We show in the following these linguistic aspects for Algiers dialect.

1) **Verbs conjugation:** Verb conjugation in ALG is affected (as in MSA) by person (first, second or third person), number(singular or plural), gender (feminine or masculine), tense (past, present or future), and voice (active or passive). Algiers dialect uses as MSA the followings forms:

- The past: Its forms are obtained by adding suffixes relative to number and gender to the verb root and by changing its diacritic marks (see Table VI for a sample).
TABLE VI: The verb كتب conjugation in the past tense.

| Pronouns | ALG | MSA | English |
|----------|-----|-----|---------|
| 1st Person | كنتَ كتب | أنا كنتَ كتب | I wrote |
|          | كنتَ كتب | حنا | We wrote |
| 2nd Person | كنتَ كتب | أنتَ كنتَ كتب | You wrote |
|          | كنتَ كتب | أنتَ | You wrote |
|          | كنتَ كتب | أنتَ أنتَ كنتَ كتب | You wrote |
| 3rd Person | كنتَ كتب | هيا | She wrote |
|          | كنتَ كتب | هو | He wrote |
|          | كنتَ كتب | هوا | They wrote |

- The present and future: The present form of a ALG verb is achieved by affixation: the prefixes تَ دَى and تَ دَى and the suffixes ى and o (Table VIII). The verb could be preceded by the particle َ ﻓ (in its inflected form). to express a present continuous tense. The future is obtained in the same way as present (same prefixes and affixes) but it must be marked by the ante-position of a particle or an expression that indicates the future like أومعد or غدوا (tomorrow), next month, ... etc.

TABLE VII: The verb لعب conjugation in the present tense.

| Pronouns | ALG | MSA | English |
|----------|-----|-----|---------|
| 1st Person | لعبتُ | أنا لعبتُ | I play |
|          | لعبتُ | حنا | We play |
| 2nd Person | لعبتُ | أنتَ لعبتُ | You play |
|          | لعبتُ | أنتَ | You play |
|          | لعبتُ | أنتَ أنتَ لعبتُ | You play |
| 3rd Person | لعبتُ | هيا | She plays |
|          | لعبتُ | هو | He plays |
|          | لعبتُ | هوا | They play |

- The imperative: It expresses commands or requests, and is used only for the second person. It is generally realised by adding the prefix ٌأ and the suffixes ى and o to the verb.

TABLE VIII: The verb خرج conjugation in the present tense.

| Pronouns | ALG | MSA | English |
|----------|-----|-----|---------|
| 1st Person | خرجتُ | أنتَ خرجتُ | Get out (you, singular, feminine) |
|          | خرجتُ | أنتَ خرجتُ | Get out (you, singular, masculine) |
| 2nd Person | خرجتُ | أنتَ أنتَ خرجتُ | Get out (you, plural, feminine & masculine) |

2) Declension: Singular word declension in written Arabic corresponds to three cases: the nominative, the genitive, and the accusative which take the short vowels ¯ َ and ُ respectively attached to the end of the word. These three cases are used to indicate grammatical functions of the words. It should be noted that also the vowels ( ٌَ ُ) represent the tanween doubled case endings corresponding to the three cases cited above and express nominal indefiniteness. ALG has dropped these case endings such as all Arabic dialects. The disappearance of final short vowels and dropping of /h/ in certain conditions in many dialects of Arabic are very significant changes [10]. The same author in [8] states: Classical Arabic has three cases in the noun marked by endings; colloquial dialects have none. Thus, a major feature of ALG is that it does not accept the three cases declension of singular nouns and adjectives as written Arabic. For singular nouns declension to the plural, ALG have the same plural classes as MSA:

- Masculine regular plural: which is formed without modifying the word structure by post-fixing the singular word by َ ﻃ ﻊ, Unlike written Arabic where the masculine regular plural of a noun is obtained by adding the suffixes َ وْن (for the nominative), and َ ﻆ ﻊ (for both the accusative and genitive) depending on the grammatical function of the word. For example, masculine regular plural of MSA word معلم (teacher) could be معلمون (nominative case) or معلمين (accusative or genitive). In contrast, for instance the ALG word رأب (going) always takes َ راًب (for the regular plural whatever its grammatical category).

- Feminine regular plural: is obtained by adding the suffix َ أٌن to the word without changing the structure of the word as in MSA but with a single difference in case endings. Indeed, in MSA, the feminine regular plural has the following marks cases (أَن or آَن for nominative and آَن or آَن for accusative and genitive), ALG has only one mark case which is the Sukun (absence of diacritic whose symbol is ). For example the plural of MSA word جيلات (girl) or جيلات and the plural of ALG word شاب (both MSA and ALG words mean beautiful).

- Broken plural: an irregular form of plural which modifies the structure of the word to get its plural. As in MSA it has different rules depending on the word pattern. Like singular words, the MSA broken plural takes the three case endings in ALG it does not.

In Table IX we give an example for each ALG plural category.

Another major difference between Algiers dialect and the written Arabic is the absence of the dual (a kind of plural which designs 2 items). Indeed in MSA, for example the dual of ولد (a boy) is designed by ولدان ( the word is post-fixed by َ ﻊ ﻆ or َ ﻉ ﻊ depending on the case). In ALG Generally, the dual is obtained by the word زوج (two) followed by the plural

8 َ جيلات or جيلات also.
9 َ آٌن for nominative case and َ ﻆ for both accusative and genitive
TABLE IX: Examples of ALG plural forms.

| Plural       | ALG           | MSA          | English                  |
|--------------|---------------|--------------|--------------------------|
| Regular masculine | فلاحيين/Faláyun | فلاحيون/Faláyun | Farmer/Farmers            |
| Case ending  | No vowel      |ينكون    | In What                  |
| Regular feminine | طيبيات/Te'ebiat | طيبيات/Te'ebiat | Doctor/Doctors            |
| Case ending  | No vowel      |ات       | With what                |
| Irregular    | طيور/Te'yur   |ييمم     | Day/Days                 |
| Case ending  | No vowel      |ييمم     |                         |

(feminine or masculine) of the noun or the adjective. For example, the dual of ولد is زوج ولد (two boys)

C. Syntactic level

1) Declarative form: Words order of a declarative sentence in ALG is relatively flexible. Indeed, in common usage ALG sentences could begin with the verb, the subject or even the object. This order is based on the importance given by the speaker to each of these entities; usually the sentence begins with the item that the speaker wishes to highlight. In Table X, we give an example of different word orders for a same sentence. It should be noted that the two first forms (SVO, VSO) are the most used in the every day conversations.

2) Interrogative form: In Algiers, any sentence can be turned into a question, in any one of the following ways:

1) It may be uttered in an interrogative tone of voice, like راح تقرأ؟ (Will you revise?).
2) By introducing an interrogative pronoun or particle as وين راح تقرأ؟ (where will you revise?).

We list in Table XI the most common interrogative particles and pronouns used in the dialect of Algiers. We mention particularly the particle أيك used in questions that accept a yes or no answer.

3) Negative form: The particles ما and ماتي are generally used to express negation. ما is used both in Algiers’s dialect and MSA, but the form of negation differs between the two languages whereas ماتي is specific to the ALG. Using these particles, the negative form is obtained in different ways in ALG (we give in Table XII some examples labeled with each enumerated case):

- Negation with ماتي particle
  1) Adding the affixes ماتي and ش to conjugated verbs (ما as prefix and ش as suffix).
  2) We can enumerate a particular case with the particle راح which is equivalent to the verb to be in present tense. The negation is obtained by adding the affixes ماتي and ش to the particle راح possibly combined with a personal pronoun.

- Negation with ما particle
  3) The particle ما can be added at the beginning of a verbal declarative sentence without modification of the sentence.
  4) The particle ما can be added at the beginning of a verbal declarative sentence by introducing the relative pronoun أيك.
  5) In the case of a nominal sentence, ما can be added at the beginning of the sentence by reversing the order of its constituents.
  6) Also ما could be added in the middle of a nominal sentence with no modification.

Table XII illustrates some examples of declarative sentences with their negations.

10 An exception is made for words like عينين (two eyes), وديين (two ears), ...
IV. CORPUS CREATION

As mentioned above, this work began from scratch. No kind of resources was available for Algiers dialect. The foundation stone of the work was a corpus that we created by transcribing conversations recorded from everyday life and also from some TV shows and movies. This transcription step required conventional writing rules to make the transcribed text homogeneous. Considering the fact that ALG is an Arabic dialect, we adopted the following writing policy: when writing a word in Algiers dialect we look if there is an Arabic word close to this dialect word, if it does exist we adopt the Arabic writing for the dialect word, otherwise the word is written as it is pronounced.

The transcription step produced a corpus of 6400 sentences that we afterwards translated to MSA. Thus, we got a parallel corpus of 6400 aligned sentences. In Table XIII, we give informations about the size of this corpus.

### Table XII: Declarative sentences with their Negation.

| Case | ALG                      | MSA                      | English          |
|------|--------------------------|--------------------------|------------------|
| 1    | ملاعبت                   | لم تلعب / ما لعبتي        | she played       |
| 2    | راهي مريضة               | أنها مريضة               | She is ill       |
| 3    | لهما كنروا                 | ليسا كنروا               | They wrote       |
| 4    | لهما كنروا                 | ليسا كنروا               | They wrote       |
| 5    | الأولد مريض               | ليس الأولد مريض          | The boy is ill   |
| 6    | الأولد مريض               | ليس الأولد مريض          | The boy is ill   |

### Table XIII: Parallel corpus description.

| Corpus | #Distinct words | #Words |
|--------|----------------|--------|
| ALG    | 8966           | 38707  |
| MSA    | 9131           | 40906  |

It should be noted that all tasks described above were done by hand. It was time consuming but the result was a clean parallel corpus. Furthermore, ALG side of this corpus has been vocalized with our diacritizer described in [11] and used to develop the first NLP resources dedicated to an Algerian dialect (at our knowledge). The next sections of this paper are dedicated to describe these resources.

V. GRAPHEME-TO-PHONEME CONVERSION

As pointed out above, the general purpose of the project TORJMAN is a speech translation system between Modern Standard Arabic and Algiers dialect. Such a system must include a Text-to-Speech module that requires a Grapheme-to-Phoneme converter. We therefore dedicated our efforts to develop this converter by using ALG vocalized corpus described earlier.

Grapheme-to-Phoneme (G2P) conversion or phonetic transcription is the process which converts a written form of a word to its pronunciation form. Grapheme phoneme conversion is not a simple deal, especially for non-transparent languages like English where a phoneme may be represented by a letter or a group of letters and vice-versa. Unlike English, Arabic is considered a transparent language, in fact the relationship between grapheme and phoneme is one to one, but note that this feature is conditioned by the presence of diacritics. Lack of vocalization generates ambiguity at all levels (lexical, syntactic and semantic) and the phonetic level consequently, such as the word /كتب/ (kataba), its phonetic transcription could be /kataba/, /kutiba/, /kutubun/, /kutub/, /katbin/... Algiers dialect obeys to the same rule, without diacritics grapheme-phoneme conversion will be a difficult issue to resolve.

Most works on G2P conversion obey to two approaches: the first one is dictionary-based approach, where a phonetized dictionary contains for each word of the language its correct pronunciation. The G2P conversion is reduced to a lookup of this dictionary. The second approach is rule-based [12], [13], [14], in which the conversion is done by applying phonetic rules, these rules are deduced from phonological and phonetic studies of the considered language or learned on a phonetized corpus using a statistical approach based on significant quantities of data [15], [16]. For Algiers dialect which is a non-resourced language, a dictionary based solution for a G2P converter is not feasible since a phonetized dictionary with a large amount of data is not available. The first intuitive approach (regards to the lack of resource) is a rule based one, but the specificity of Algiers dialect (that we will detail hereafter in the next section,) had led us to a statistical approach in order to consider all features related to this language.

A. Issues of G2P conversion for Algiers dialect

Algiers dialect G2P conversion obeys to the same rules as MSA. Indeed, ALG could be considered as a transparent language since alignment between grapheme and phoneme is one to one when the input text is vocalized. But unfortunately, it is not as simple as what has been presented, since ALG contains several borrowed words from foreign languages which most of them have been altered phonologically and adapted to it. Henceforth, the vocabulary of this dialect contains many French words used in everyday conversation. French borrowed words could be divided into two categories: the first includes French words phonologically altered such as the word /سورة/ (sûr in French, sure) whose utterance is /sûra/ /sûr/ is not an Arabic phoneme but a French phoneme). This last category constitutes...
a serious deal for G2P conversion since these words do not obey to Arabic pronunciation rules.

| Dialect word | Dialect phonetic transcription | French word | English |
|--------------|------------------------------|-------------|---------|
| كوزينا        | /ku:zinə/                    | Cuisine     | Kitchen |
| تابل          | /ta:bla/                     | Table       | Table   |
| كونرسالون     | /konrerksalən/              | Connection  | Connexion |
| دوزير        | /duzir/                      | Devise      | Currency |

In the examples of Table XIV although the first two words are French, they are phonetized as Arabic words. The French phoneme /t/ is replaced by the Arabic Phoneme /t'/ in the word table. On the other side, the last two words are phonetized as French words since they are pronounced as in French by Algiers dialect speakers. In order to take account of this word category, the French phonemes like /t/, /s/ and /l/ must be included in Algiers dialect phonemes.

**B. Rule based approach**

As stated previously, the rule based approach for G2P conversion applied to ALG requires a diacriticized text, that is why we used our ALG vocalized corpus. The diacritized text is converted into its phonetic form by applying the followings rules. It should be noted that most of these rules are those adopted also for Arabic [12], [13] and are applicable only for Arabic words and foreign words phonologically altered in our corpus.

Let consider: BS is a mark of the beginning of a sentence, ES is a mark of the end of a sentence, BL is a blank character, SC is a consonant, V is a vowel, LC is a lunar consonant, and LV is a long vowel. A sample conversion rule could be written as follows:

\[
LFT + GR + RGT \Rightarrow /PH/ 
\]

The rule is read as follows: a grapheme GR having as left and right contexts LFT and RGT respectively, is converted to the phoneme PH. Left and right contexts could be a grapheme, a word separator, the beginning or the end of a sentence or empty.

We give in the following all rules that we used for Algiers dialect G2P (the representation of these rules according to the sample below is given in the Appendix (Table XXIII).

1) د , ط , ط و ث , ث rules

In Algiers dialect, the letters د , ط , ط and ث , ث are not used, they are in most cases pronounced as the graphemes ة , ض and َ، ث , respectively.

2) Foreign letters rules Algiers dialect alphabet corresponds to Arabic alphabet extended to three foreign letters G, V and P.

3) Definite article ال

- The definite article ال is not pronounced when it is followed by a lunar consonant (which does not assimilate the ـ).
  
  Example: ال القمر (the moon) ➞ /laqmar/

  This rule is the same as in MSA with the difference that in MSA the ـ is pronounced if the definite article is in the beginning of the sentence.

- When the definite article ال is followed by a solar consonant the ل is not pronounced and the consonant following the ل is doubled (gemination).

  Example: السقف (the roof) ➞ /?assqal/

- When the definite article ال is preceded by a long vowel ـ and followed by a solar consonant the definite article is omitted and the solar consonant is doubled (gemination).

  Example: فدار ➞ /fddAr/

4) Words Case-ending

Words case ending in Algiers dialect is the Sukun (Absence of diacritics), so the last consonant of a word should be pronounced without any diacritic.

Example : قبل (before) ➞ /qbal/

5) Long vowel rules

When and ى appear in a word preceded by the short vowels ـ , ـ and ـ , respectively, their relative long vowels are generated.

Examples:

- كأس (a cup) ➞ /kaːs/
- فول (beans) ➞ /fuːl/
- كيبور (a well) ➞ /kbiːt/

6) Glottal stop rule

In Algiers dialect, when a word begins with a Hamza, its phonetic representation begins with a glottal stop. In the end of a word the Hamza preceded by ـ is not pronounced.

Example: السمكت (stop talking) ➞ /ʔaskut/ and (sky) ➞ /smʔ/.

It should be noted that the Hamza in the middle of the word is replaced by the long vowels ـ or ى in Algiers dialect. For example the Arabic words بئر (hole) and فأس (poleax) correspond to /bəɾ/ and /faːs/ respectively.

7) Alif Maqṣura rule

Alif Maqṣura (ى) (which is always preceded by a fatha) at the end of a word is realized as the short vowel /a/.

Example: رمي (he throws) ➞ /ɾmaː/.

8) Alif Madda ً

Alif madda ً is realized as alef /ʔ/ with the long vowel /aː/.

Example: آمن (he trusts) ➞ /ʔaːmən/.

9) Words ending with ـ

The ـ is not pronounced in Algiers dialect unlike in MSA where it is realized with the two phonemes /ʔ/ and /u/ (depending on the word position).

Example: طالعة (a girl) ➞ /ʔ’afla/.

10) Words ending with ـ
The ş is not pronounced in Algiers dialect when it is preceded by ل. Example: كتاب (his book)⇒ /kta:bu/

11) Words containing the sequences ب. ن When a ن is followed by a ب, the ن is pronounced as /m/
Example: مبر (a foretop) ⇒ /mambar/
12) Gemination rule When the Shadda appears on a consonant, this consonant is doubled (geminated)
Example: سكر (sugar) ⇒ /sukkur/
It should be noted that most of these rules could be applied for other Algerian dialects and Arabic dialect close to them such Tunisian and Moroccan.

Experiment: As indicated above for experiment we used our ALG vocalized corpus which includes three categories of words:

1) Arabic words.
2) French words phonologically altered and their pronunciation is realized with Arabic phonemes.
3) French words for which the pronunciation is realized with French phonemes.

We applied phonetization rules seen below on the ALG corpus. In addition to Arabic words, French words of the second category are correctly phonetized because their phonetic realization is close to Algiers dialect. For example the word کوزينة (kitchen, original French word cuisine) which is a borrowed French word phonologically altered is correctly converted as /kuzina/, while a word in the third category as کونکسيون (connexion, original French word connexion) is incorrectly converted to /kunksjou:n/ since it is realized /kɔnkrksjøː/ with French phonemes. Considering these words, system accuracy is 92%. The issue of these words is that we can not introduce rules for French words written in Arabic script, since the relation between Arabic graphemes and French phonemes is not one to one. For example the graphemes ـ in a French word written in Arabic script could correspond to the French phonemes /l/, /l/ or /O/ (see some examples in Table XV).

| Dialect word | French phonetic transcription | French word | English |
|--------------|------------------------------|-------------|---------|
| سور         | syn                          | Sure        | Sure    |
| بور         | p.m                         | Port        | Port    |
| سودوز       | sudex                       | Soudre      | Wilder  |

TABLE XV: Examples of mappings between Arabic grapheme ـ and French phonemes.

C. Statistical Approach

Rule based approach adopted above does not take into account French words used in ALG which are pronounced as in French language. This issue takes us to choose a statistical approach in order to consider this feature. We use statistical machine translation system where source language is a text (a set of graphemes) and target language is its phonetic representation (a set of phonemes). This system uses Moses package[17], Giza++[18] for alignment and SRILM[19] for language model training. The main motivation of using a statistical approach is that we can include French phonemes in the training data. For building this system, the first component is a parallel corpus including a text and its phonetic representation. Actually, this resource is not available, so we created it by using the rule based converter described above. We proceed as follows: we used the rule based system to convert Arabic words and French words phonologically altered (category 1 and 2) to Arabic phonemes. Whereas for French words realized with French phonemes (category 3), we began by identifying them and we transliterated them to their original form in Latin script, then converted them to French phonemes (using a free French G2P converter), all these operations were done by hand. For example the word كونكسيون is transliterated to connexion then converted to /kɔnkrksjøː/.

This system operates at grapheme and phoneme level, we split the parallel corpus into individual graphemes and phonemes including a special character as word separator in order to restore the word after conversion process (see Table XVI).

TABLE XVI: Examples of aligned graphemes and phonemes.

| Arabic | French | English |
|--------|--------|---------|
| Null   | /o     | o       |
| /u     | /u      | u       |
| /e     | /e      | e       |

Experiment: For evaluating the statistical approach, we split the parallel corpus into three datasets: training data (80%) tuning data (10%) and testing data (10%). First we tested the statistical approach on a corpus containing only Arabic words and French words phonologically altered (category 1 and 2). We got an accuracy of 93%. Then we proceeded to a test on a corpus including the three words categories, system accuracy decreases to 85%. This result is due to the increase of hypothesis number of each grapheme because of introducing French phonemes in the training data. The graphemes ـ for example in some Arabic words (category 1) are phonetized as the French phonemes /l/ or /l/ instead of the Arabic long vowel /u:/, the phoneme /s/ instead of /u:n/. Contrary to that some words in category 3 are phonetized with Arabic phonemes by substituting for example the phonemes /l/, /l/ or /O/ by the /l/, and /s/ by /s/.

D. Discussion

At first glance, and regards to accuracy rates, we could deduce that rule based approach is more efficient than statistical approach (92% vs 85%). Rule based approach does not take into account French words of category 3, it achieves efficient results only for Arabic words and French phonologically altered words (category 1 and 2). Results of statistical approach must be analysed regards to the small amount of the training data. On another side, a hybrid approach could be adopted: instead of using one corpus including all categories of words for training the statistical G2P converter, we can use two corpora: the first one including words of categories 1 and 2, could be processed by rule based approach. The second corpus is a parallel corpus including words of category 3 with their French phonetization used for training the statistical G2P
VI. MORPHOLOGICAL ANALYZER FOR ALGERIAN DIALECT

A. Related works

Compared to MSA, there are a little number of Morphological Analysers (MA) dedicated to Arabic dialects. Works in this area could be divided into two categories. The first one includes MA that are built from scratch such as in [20] and [21], the second includes works that attempt to adapt existing MSA Morphological Analysers to Arabic dialect. This trend is adopted for several dialects since it is not time consuming. In [22], authors used BAMA Buckwalter Arabic Morphological Analyser [3] by extending its affixes table with Levantine/Egyptian dialectal affixes. The same approach is adopted in [23] where a list of dialectal affixes (belonging to four Arabic dialects) was added to Al-Khalil [24] affix list. Authors in [25] converted the ECAL (Egyptian Colloquial Arabic Lexicon) to SAMA (Standard Modern Arabic Analyser) representation [26]. For Tunisian dialect, authors in [27] adapted Al-Khalil MA, they create a lexicon by converting MSA patterns to Tunisian dialect patterns and then extracting specific roots and patterns from a training corpus that they created.

B. Adopted Approach

To build a MA for Algiers dialect, we decide to adapt BAMA, since it does not consume time and takes profit from the fact that it is widely used. BAMA is based on a dictionary of three tables containing Arabic stems, suffixes and prefixes and three compatibility tables defining relations between stems, prefixes and suffixes. Adaptation of BAMA is got by populating these tables by dialect data.

C. Building the dialect dictionary

We built dialect dictionary by adopting the following principle: in order to exploit BAMA dictionary, we kept from it all entries that belong also to ALG with some modification (for example MSA prefixes and are used in ALG so we kept them as ALG prefixes). Beside that, we deleted all entries which are not suitable for Algiers dialect. Moreover, we created entries that are purely dialectal and which did not exist in MSA dictionary.

1) Affixes tables: For affixes tables, common affixes between MSA and ALG are kept (in prefixes and suffixes tables), whereas all other MSA affixes which do not belong to dialect were deleted. However, some dialect affixes which do not exist in MSA were added to affixes tables. Note that when an affix is deleted, all complex affixes where it occurs are also deleted.

1) Prefixes table: We kept some prefixes unchanged like prefixes and that precede imperfect verbs (for the singular third person masculine and feminine, respectively). We eliminated purely MSA prefixes.

| Kept pref. | Description |
|-----------|-------------|
| ي، ۚ | Imperfect Verb Prefix (sing., third person, masc., fem.) |
| ل، ۚ | Noun Prefix (definite article) |
| ب، ۚ | Preposition Prefix |

Del. pref. | Description |
|-----------|-------------|
| ف | Conjunction Prefix |
| س | Future Imperfect Verb Prefix |
| د | Conj. Pre.+Preposition Pre.+Definite Art. Pre. |

Add. pref. | Description |
|-----------|-------------|
| ت، ۚ | Perfect verb pre. (past voice, (sing., masc.) and (plu, masc/fem.)) |
| ت، ۚ | Perfect verb pre. (past voice, (sing. fem.)) |

2) Suffixes table: We also eliminated all MSA suffixes not used in Algiers dialect mainly:
- Suffixes related to the dual both feminine and masculine,
- Feminine plural suffixes,
- All word case endings suffixes

All complex suffixes where they appear were also deleted. Likewise, we added dialectal suffixes like the suffix (expressing the future when it precedes imperfect verbs) and the prefix (conjunction), some examples are given in Table XVII.

| Kept pref. | Description |
|-----------|-------------|
| ي، ۚ | Imperfect Verb Prefix (sing., third person, masc., fem.) |
| ل، ۚ | Noun Prefix (definite article) |
| ب، ۚ | Preposition Prefix |

Del. pref. | Description |
|-----------|-------------|
| ف | Conjunction Prefix |
| س | Future Imperfect Verb Prefix |
| د | Conj. Pre.+Preposition Pre.+Definite Art. Pre. |

Add. pref. | Description |
|-----------|-------------|
| ت، ۚ | Perfect verb pre. (past voice, (sing., masc.) and (plu, masc/fem.)) |
| ت، ۚ | Perfect verb pre. (past voice, (sing. fem.)) |

2) Stems table: Dialect stems table was populated by the lexicon of Algiers dialect corpus and MSA stems included in BAMA. We used a part (85%, 9170 distinct words) of our ALG corpus for creating dialect stems, the remaining 15% (1618 distinct words) is used for test.

Stems from ALG corpus lexicon

First, we began by extracting a list of nouns easily identifiable by affixes and definite article (used only with nouns). We deleted these two affixes from all extracted words, then from obtained list of words we created stems entries according to BAMA. Next, the rest of the corpus was analysed and classified into three sets: function words, verbs and nouns (which do not include and suffixes) and converted to stems according to BAMA stems categories. Let us indicate that we added some stems categories to take into account all dialectal features. For example, in MSA the perfect verb stem category

12Prefixes that could not belong to Algiers dialect.

13Note that as MSA conjunction prefix has been deleted (since it does not exist in ALG), and as preposition prefix has been created.
TABLE XVIII: Examples of kept, deleted and added suffixes in ALG suffixes table.

| Kept Aff. | Description                     |
|-----------|---------------------------------|
| åK        | Accusative/genitive noun Suffix(masc.,plu.) |
| H@        | Noun Suffix(fem.,plu.)          |
| H         | perfect verb suffix (fem.,sing) |

| Del. suff. | Description                     |
|------------|---------------------------------|
| ä          | Perfect/Imperfect Verb Suffix(subject, plu., fem.) |
| ÁÖ ß        | Perfect/Imperfect Verb Suffix(subject, dual., fem/masc., 2nd person) |
| AÒë        | Perfect/Imperfect Verb Suffix(direct object, dual., fem/masc., 3rd person) |
| àð         | Nominative Noun Suffix (masc.,plu.) |
| à@         | Nominative Noun Suffix (masc.,dual) |
| áë         | Perfect/Imperfect Verb Suffix(direct object, plu., fem.) |
| áêK        | Perfect Verb Suffix(subject sing.,2nd person,masc.,direct object, plu., 3rd person, fem.) |

| Add. suff. | Description                     |
|------------|---------------------------------|
| €          | Perfect/Imperfect Verb Negation Suffix |
| Òë        | Perfect/Imperfect Verb Negation Suffix (direct object,plu., 3rd person, masc./fem) |
| Ò»        | Perfect/Imperfect Verb Negation Suffix (direct object,plu., 2nd person, masc./fem.) |
| ð          | Per./Imp. Verb Suffix(direct object,plu.,masc./fem.) |

with the pattern ٌفعل يُقَلُ covers the three persons, the two genders, the single, the dual and plural; just relative suffixes are added to it to have its different inflected forms. In ALG, we split this stem category into two distinct stems: ٌفعل and ٌفعل يُقَلُ to cover all perfect verbs inflected forms, in Table XIX we give an example related to the stem ٌفعل يُقَلُ (to hear).

TABLE XIX: Example of splitting a MSA stem to two Dialectal stems.

| Eng. pro. | Dia. pro. | Dia. verb | Dia. stem | MSA pro. | MSA verb | MSA stem |
|-----------|-----------|-----------|-----------|----------|----------|----------|
| She       | हे        | नहीं है    | हे है    | हे है    | हे है    | हे है    |
| They      | वे        | नहीं है    | वे है    | वे है    | वे है    | वे है    |
| He        | हे        | नहीं है    | हे है    | हे है    | हे है    | हे है    |
| We        | हम        | नहीं है    | हम है    | हम है    | हम है    | हम है    |

Exploiting MSA BAMA stems

1) **Verbs**

The main idea for creating ALG verb stems from MSA stems is using verbs pattern. For example the verbs having ALG pattern ٌفعل يُقَلُ or ٌفعل يُقَلُ are in most cases Arabic verbs with the pattern ٌفعل يُقَلُ or ٌفعل يُقَلُ. Some other ALG verbs keep the same pattern as in MSA like verbs with the patterns ٌفعل يُقَلُ.

From stems table, we extracted all perfect verbs having the patterns ٌفعل يُقَلُ and ٌفعل يُقَلُ. After that, the verbs having the three first patterns are converted to Algiers dialect pattern by changing diacritic marks to ٌفعل while the verbs corresponding to pattern ٌفعل are kept as they are (since this pattern is used in ALG). At this stage, we constructed a set of Arabic verb stems having dialect pattern, we analysed them and eliminated all stems that are not used in ALG. We give in Table XXI some examples.

TABLE XX: Examples of converted stems from MSA to ALG.

| Stems | ALG Dialect | MSA | English |
|-------|-------------|-----|---------|
| हे है    | हे है    | हे है    | हे है    |
| वे है    | वे है    | वे है    | वे है    |
| हे है    | हे है    | हे है    | हे है    |
| हम है    | हम है    | हम है    | हम है    |

2) **Nouns**

We kept all proper nouns from MSA stems table because it contains an important number of entries related to countries, currencies, personal nouns,... We analysed all other types of words and kept from them those existing in ALG by modifying diacritics, adding or deleting one or more letters.

3) **Function words**

We deleted all function words that do not exist in ALG like relative pronouns and personal pronouns related to the dual and feminine plural, then we translated remaining ones to ALG.

Note that we introduced dialect stems with non Arabic letters ٍ G, ٍ V, and ٍ P in stems table and we modified BAMA code to consider words containing these letters. Also, since every stem entry in BAMA contains an English glossary, when creating a dialect entry, we added the Arabic word to English glossary, so for each dialect entry is associated an English and Arabic glossary.

After creating affixes and stems tables for ALG, compatibility tables of BAMA were updated according to the data included in these tables.

D. **Experiment**

As mentioned above, we tested our MA on the Algiers Dialect corpus, the test set contains 1618 distinct words extracted from 600 sentences chosen randomly. We consider
that a word is correctly analysed if it is correctly decomposed to prefix+stem+suffix and if all the features related to them are correct (POS, gender, number, person). We first began by testing the MA with stems extracted only from the ALG corpus lexicon, then we introduced stems created from the MSA stems table. We list in Table XXI the obtained results.

**TABLE XXI: Results of ALG morphological Analyser.**

| Results | ALG corpus stems | MSA stems+ALG corpus stems |
|---------|------------------|---------------------------|
| # Analysed words | 703 | 1115 |
| Percentage | 43% | 69% |
| # Unanalysed words | 915 | 503 |
| Percentage | 57% | 31% |

We examined the words for which no answer were given by the morphological analyzer (see Table XXII), most of the cases are:

- French words which do not exist in the stem table like أكبيروس (´electricité, electric), or words like النمو (´numéro, number) that are included in stems table but with an other orthography (respectively النمو and أكبيروس). The same case is observed for nouns written with long vowel in the end instead of ُ(place).
- We noticed also that some words are written with missed letters like the word أنس (´inside) which appears in stems table as أنساء. The same case is noticed for قال (he said to me) instead of قال or قال (I said to him) instead of قال or قال.
- Some Unanalysed words also are proper nouns.

**TABLE XXII: Examples of unanalysed words.**

| Unanalysed word | Corresponding stem | English |
|------------------|--------------------|---------|
| أكبيروس | Internet |
| أممم | After |
| أكبيروس | Trimester |
| تليفون | Phone |

**REFERENCES**

[1] S. Harrat, K. Mefteh, M. Abbas, and K. Smaili, “Building resources for algerian arabic dialects,” in Proceedings of Interspeech, 2014, pp. 2123–2127.

[2] ——, “Grapheme to phoneme conversion: An arabic dialect case,” in Proceedings of 4th International Workshop On Spoken Language Technologies For Under-resourced Languages SLTU, 2014, pp. 257–262.

[3] B. Tim, “Buckwalter arabic morphological analyzer version 1.0,” Linguistic Data Consortium LDC2002L49, 2002.

[4] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schone, R. Schwartz, and D. Vergyri, “Novel approaches to arabic speech recognition: Report from the 2002 johns-hopkins summer workshop,” in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP ’03), vol. 1, April 2003, pp. I–344–I–347.

[5] R. Hetzron, *The Semitic Languages*, ser. Routledge language family descriptions. Routledge, 1997. [Online]. Available: https://books.google.com/books?id=nbUOAAAAQAAJ

[6] J. C. Watson, *The phonology and morphology of Arabic*. Oxford university press, 2007.

[7] A. Boucherit, *L’Arabe parlé à Alger*. ANEP Edition, 2002.

[8] C. A. Ferguson, “Diglossia,” Word, vol. 15, pp. 325–340, 1959.

[9] F. H. Amr, B. A. Adaileh, and B. A. Rakhieh, “Arabic diglossia: A phonological study,” Argumentum 7, Debreceni Egyetemi Kiado, Tanulmány, pp. 19–36, 2011.

[10] C. A. Ferguson, “Two problems in arabic phonology,” Word, vol. 13, pp. 460–478, 1957.

[11] S. Harrat, M. Abbas, K. Mefteh, and K. Smaili, “Dicritications for arabic dialect texts,” in Proceedings of Interspeech, 2013, pp. 125–132.

[12] M. Alghamdi, H. Almuhtasab, and M. Alshifi, “Arabic phonological rules,” *Journal of King Saud University: Computer Sciences and Information (in Arabic)*, vol. 16, pp. 1–25, 2004.

[13] Y. A. El-Imam, “Phonetization of arabic: rules and algorithms,” *Computer Speech Language*, vol. 18, no. 4, pp. 339–373, 2004.

[14] M. Zeki, O. O. Khalifa, and A. Naji, “Development of an arabic text-to-speech system,” in *International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2010, pp. 1–5.

[15] P. Taylor, “Hidden markov model for grapheme to phoneme conversion,” in *Proceedings of Interspeech*, 2005, pp. 1973–1976.

[16] K. U. Ogureke, P. Cahill, and J. Carson-Berndsen, “Hidden markov models with context-sensitive observations for grapheme-to-phoneme conversion,” in *Proceedings of Interspeech*, 2010, pp. 1105–1108.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, demonstration session, pp. 177–180, 2007.
[18] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, volume 29, number 1, pp. 19–51, 2003.

[19] A. Stolcke, “Srilm – an Extensible Language Modeling Toolkit,” in *Proceedings of Interspeech*, Denver, USA, 2002, pp. 901–904.

[20] N. Habash and O. Rambow, “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*. Association for Computational Linguistics, 2006, pp. 681–688.

[21] M. Altantawy, N. Habash, and O. Rambow, “Fast yet rich morphological analysis,” in *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 116–124.

[22] W. Salloum and N. Habash, “Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation,” in *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Association for Computational Linguistics, 2011, pp. 10–21.

[23] K. Almeman and M. Lee, “Towards developing a multi-dialect morphological analyzer for arabic,” in *4th International Conference on Arabic Language Processing*, 2012, pp. 19–25.

[24] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. A. O. Bebah, and M. Shoul, “Alkhailil morpho sys: A morphosyntactic analysis system for arabic texts,” in *Proceedings of 7th International Computing Conference in Arab ACIT*, 2011.

[25] N. Habash, R. Eskander, and A. Hawwari, “Morphological analyzer for egyptian arabic,” in *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON*. Association for Computational Linguistics, 2012, pp. 1–9.

[26] D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter, “Standard arabic morphological analyzer (SAMA) version 3.1,” *Linguistic Data Consortium LDC2009E73*, 2009.

[27] I. Zribi, M. E. Khemakhem, and L. H. Belguith, “Morphological analysis of tunisian dialect,” in *International Joint Conference on Natural Language Processing*, 2013, pp. 992–996.
TABLE XXIII: Algiers dialect Rules for G2P conversion.

| #  | Rule title                                      | Rule                                      |
|----|------------------------------------------------|-------------------------------------------|
| 1  | TD, TDh, and TDh rules                         | $\{C, V\} + \ddot{d} + \{C, V\} \Rightarrow /d/ $<br> $\{C, V\} + \ddot{v} + \{C, V\} \Rightarrow /d/ $<br> $\{C, V\} + \ddot{v} + \{C, V\} \Rightarrow /T/ $ |
| 2  | Foreign letters rules                          | $\{C, V\} + \ddot{g} + \{C, V\} \Rightarrow /g/ $<br> $\{C, V\} + \ddot{h} + \{C, V\} \Rightarrow /v/ $<br> $\{C, V\} + \ddot{y} + \{C, V\} \Rightarrow /p/ $ |
| 3  | Definite article حَدَّسَةٌ $\{LC\} + {\ddot{g}} + \{BL, BS\} \Rightarrow /l/ + /LC/ $<br> $\{SC\} + {\ddot{g}} + \{BL - BS\} \Rightarrow /a/ + /SC/ + /SC/ $ |
| 4  | Words Case-ending                              | $\{BL, ES\} + C + \{C, V\} \Rightarrow /C/ $ |
| 5  | Long vowel rules                               | $\{C+\ddot{a}\} + \ddot{a} + \{C\} \Rightarrow /a : / $<br> $\{C+\ddot{u}\} + \ddot{u} + \{C\} \Rightarrow /u : / $<br> $\{C+\ddot{i}\} + \ddot{i} + \{C\} \Rightarrow /i : / $ |
| 6  | Glottal stop rule                              | $\{C, V\} + \ddot{?} + \{BS, BL\} \Rightarrow /?/ $<br> $\{BL, ES\} + \ddot{?} + \{1\} \Rightarrow /Null/ $ |
| 7  | Alif Maqṣura rule $\ddot{y} \{BL, ES\} + \ddot{y} + \{C\} \Rightarrow /a/ $ |
| 8  | Alif Madda $\ddot{a} \{C\} + \ddot{a} + \{C\} \Rightarrow /?a : / $ |
| 9  | Words ending with $\ddot{a} \{BL, ES\} + \ddot{a} + \{C, V\} \Rightarrow /Null/ $ |
| 10 | Words ending with $\ddot{a} \{BL, ES\} + \ddot{a} + \{C\} \Rightarrow /Null/ $ |
| 11 | Words containing the sequences $\ddot{b} \{b\} + \ddot{a} + \{C, V\} \Rightarrow /m/ $ |
| 12 | Gemination rule                                | $\{V\} + \ddot{w} + \{C\} \Rightarrow /CC/ $ |