# Morphological Analyzers of Arabic Dialects: A survey

Ridouane Tachicart, Karim Bouzoubaa, Salima Harrat, Kamel Smaïli

# Morphological Analyzers of Arabic Dialects: A survey

**Ridouane Tachicart[1], Karim Bouzoubaa[1], Salima Harrat[2] and Kamel Smaïli[3]**

1: Université Mohammed V (Maroc)

2: Ecole Nationale Informatique (Algérie)

3 : Université Lorraine- Loria (France)

ridouane.tachicart@research.emi.ac.ma, bouzoubaa@emi.ac.ma, lgsmus@yahoo.fr, smaili@loria.fr

***Abstract*— Morphological analysis is a crucial stage in natural language processing. For the Arabic language many attempts have been conducted to build morphological analyzers. Despite the increasing attention paid to Arabic dialects recently, only a few number of morphological analyzers have been built compared to MSA. In addition, those tools often cover a few dialects of Arabic such as Egyptian and Levantine, thereby they don't currently support all Arabic dialects. In this paper, we present a wide literature review of morphological analyzers processing Arabic dialects. We classify their building approaches and propose some guidelines to adapt them to a specific Arabic dialect. In addition, a quick benchmarking of the available analyzers is given in order to evaluate their performance. Our goal in this paper is to provide a quick reference guide of Arabic dialect morphological analyzers, as well as some recommendations for researchers needing to develop new Arabic dialect morphological analyzers. Results of this survey can be used as baseline for future Arabic dialect morphological analyzers building.**

***Index Terms*— Morphological analyzer, Arabic dialect, corpus, lexicon, natural language processing, language model, standard Arabic, benchmark.**

## I. Introduction

Arabic is a Semitic language spoken by more than 420 million people in the world. It is the official language of the Arab nation (about 22 countries) and displays a collection of forms:

• The primary form is defined as classical Arabic (known also as Quranic Arabic) found in the Quran and Jahillyah[1] literature (Arabic period before the arrival of Islam). It dates back to the 7th and 9th century from Umayyad and Abbasid times[2] where it is used in literary texts.

• Modern Standard Arabic (MSA) is the second form of Arabic currently used in formal situations such as education, government documents, broadcast news, etc. It has a strong presence in written Arabic texts since it is a high variety of Arabic with its normalization and standardization.

• Arabic dialects (AD) set is the third form of Arabic considered as the mother tongue of Arabic people. They differ from each other and are usually used in informal venues such as daily communication, TV series and programs, commercial advertising, etc. Unlike MSA, Arabic dialects have no written standards and are sparsely represented in written texts compared to MSA in spite of their recent use in internet. Arabic dialects can be divided to east Arabic and Maghreb Arabic dialects (Figure 1) according to their similarities. The first category includes Gulf (GLF), Egyptian (EGY), Levantine (LEV), Iraqi and Yemeni. While the second gathers Moroccan, Algerian, Tunisian, Libyan and Mauritanian. Many works[3] showed that the dialects of the same geographical area (eastern or western) are close. For example, Moroccan people can understand Tunisian and Algerian people better than Egyptian or Syrian people. Note that in the same country, we can divide its dialect to sub-dialects according to geographical regions.Nowadays, Arabic dialects are widely used in internet. As an illustration, Arabic people increasingly use their own dialects in social networks by expressing their opinions with these dialects. In Morocco for example, "goud.ma" and "lsvbdarija.com" are examples of websites where the text is completely written in local dialect. This is why there is an increasing interest to process Arabic dialects and build their corresponding NLP tools such as automatic identification, opinion mining and machine translation.In effect, the building of these tools relies mainly on the availability of a corresponding morphological analyzer (MA). However, there is a general lack of resources for most of these dialects. As a result, there is a slow progress in building corresponding MA and consequently a slow progress in building advanced dialect NLP tools. The availability, hence, of these tools is still in earlier stages.

To overcome this situation, it is of great importance that researchers put their efforts in building and providing, on one hand, dialect resources and, on the other hand, MA of all Arabic dialects. This is why some researchers devoted their efforts currently in assembling all the available AD resources as in the work of (Shoufan & Alameri, 2015), (Jarrar M. , Habash, Alrimawi, Akra, & Zalmout, 2016) and (Mona Diab, 2010).

Concerning the building of AD MA, there have been some attempts and our literature review revealed that researchers usually follow two directions to build them. One direction suggests to build them from scratch. Whereas, the second proposes to adapt existing MSA morphological analyzers to Arabic dialects.

[1] https://en.wikipedia.org/wiki/Jahiliyyah

[2] https://en.wikipedia.org/wiki/Classical_Arabic

[3] https://en.wikipedia.org/wiki/Variet

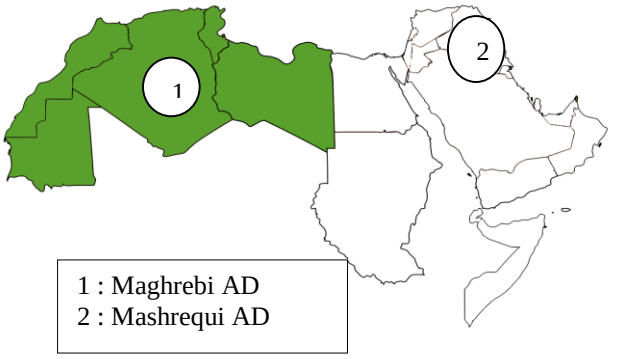


Fig.1 Arabic dialects categories according to regions

The important amount of efforts made for processing MSA led its morphological analyzers to reach high performance compared to AD analyzers. The reached accuracy exceeds 95% such as AlKhalil MA (Boudchiche, Mazroui, Ould Abdallahi Ould Bebah, & Lakhouaja, 2014). This is to be expected since MSA presents a unique language and it is shared between all Arabic people, whereas Arabic dialects are multiple and differ from one Arabic country to another. Moreover, available AD analyzers are not well exploited in advanced AD NLP applications to the best of our knowledge. As an example, they are currently only used in Machine translation or automatic language identification systems such as the work of (Salloum & Habash, Elissa: A Dialectal to Standard Arabic Machine Translation System, 2012). This problematic situation of AD analyzers raises some questions and need explanations about their:

**i) difference in reached performance**: MSA analyzers perform better than AD MA ones. Moreover, there is a large difference between AD MA reached accuracies. In fact, a set of AD analyzers reached acceptable accuracies, whereas another set reached low accuracies;

**ii) coverage of dialects**: existing AD MA miss addressing some Arabic dialects. In addition, we can find different MA for the same dialect, whereas some other dialects are only addressed by one MA;

**iii) building approaches orientation:** researchers applied different approaches in order to build AD MA. However, we notice a difference in achieved accuracies when evaluating these analyzers.

**iv) integration in NLP applications**: To the best of our knowledge, these AD MA are not used in large scale NLP systems except machine translation and automatic language identification.

The present study sheds light on Arabic dialect morphological analyzers (MA) and tries to give responses to the above questions. In fact, we present a wide literature review of Arabic dialect morphological analyzers and describe their characteristics. The study highlights also the comparison result performed on the available analyzers regarding a sample of an annotated multi-dialect corpus. Our main goal in this paper is to pave the way for researchers to select the best option for building Arabic dialect morphological analyzer or select the one that best suits their needs.

The remainder of this paper is organized as follows: Section 2 gives an overview about challenges that faces the building of AD MA and presents taken solutions. Section 3 presents related works in the field of Arabic dialect morphological analyzers. Section 4 describes followed approaches to build such morphological analyzers. In Section 5, we detail the benchmark that we have performed on existing AD MA. Then, in section 6 we provide a discussion about problematic questions; finally, we conclude the paper in section 7 with some observations.

## II. AD MA Challenges and solutions

### 1) *Challenges*

Morphological analysis of Arabic Dialects is relatively a recent area of research that gained progressive attention during the last decade. Building new Morphological analyzers for these dialects is not easy for many reasons such as:

**i) Varieties of Arabic dialects:** A set of Arabic dialects exist with linguistic differences on different levels. Moreover, each Arabic dialect displays a set of sub-dialects spoken according to regions. This situation means that a given dialect can slightly varies from a region to another in each Arabic country and thus different forms for the same dialect are spoken in. These forms of dialect differ from each other especially at the lexical and the phonological level.

**ii) Using Arabic and Latin letters (Arabizi):** Since Arabic dialects are known as spoken languages and have no standards, Arabic speakers use either Arabic or Latin letters in order to write their local dialects. Moreover, they often use Latin letters in social media as well as online chat and Short Messaging System (SMS) (Bies, et al., 2014) and thus generating massive amounts of Arabizi every day. However, current AD MA cannot process this type of text because they consider only Arabic text with Arabic letters.

**iii) Orthographic ambiguity:** AD have no standards where spelling inconsistency is a big challenge. Either using Arabic or Latin letters, the same lemma may be written in different forms according to users. For example, in Maghreb Arabic dialects, the word **بقرة** /cow/ may be also written as **بكرة** according to speakers**.**

**iv) Lack of AD resources:** Building AD morphological analyzers needs linguistic resources such as corpora and lexicons that lack currently. Only few resources were built and targeted few Arabic dialects. Moreover, they are not available in the majority of cases.

**v) Code-switching:** Typically, native speakers of Arabic tend to use a mixture of MSA and AD (when using Arabic script) in the same context especially in social media. This situation increases AD MA error rate when analyzing such text because of the MSA content.

### 2) *Proposed Solutions*

There have many efforts performed in order to overcome existing challenges and pave the way for the AD MA building such as:

**i) Focusing on main Arabic dialects:** Since several Arabic dialects are spoken in the same Arabic country according to regions, researchers proposed to focus on the main sub-dialect in each country and then deal with other sub-dialects.

**ii) Using transliteration systems:** to avoid the problem of Arabizi, several works introduced a transliteration module in their systems like the work of (May, Benjira, & Echihabi, 2014) and (Authore, 5-6 November 2017) that convert Arabic dialect

written in Latin letters into Arabic letters and then perform related processing. As a consequence, Arabizi may be handled like Arabic text in the same MA.

**iii) Conventional Orthography:** Because Arabic dialects have no standards which impede their processing, several works proposed to adopt new rules towards the standardization of AD orthography. One example is the work of (Habash, Diab, & Rabmow, Conventional Orthography for Dialectal Arabic, 2012) where they proposed a unified framework to write all AD with Arabic script based on MSA-AD similarities.

**iv) Building new AD resources:** To overcome the problem of AD resources lack, several works started from scratch and used web mining or exploited the similarities existing between MSA and AD (Shoufan & Alameri, 2015) in order to build new AD resources.

**v) Using Language Identification (LID) Systems:** The existing of code-switching in Arabic text is a major issue that increases AD MA error rate. To address this issue, several works introduced LID systems to distinguish between MSA and AD content and thus consider only AD text in processing such as AIDA system (Elfardy, Al-Badrashiny, & Diab, 2014) and the work of (Tachicart R. , Bouzoubaa, Aouragh, & Jaafar, 2017) .

## III. Related works

Many attempts for AD MA have targeted more than one Arabic dialect, while other dialects are low addressed or not concerned. In addition, we can find mono-dialect and multi-dialect morphological analyzers. In the following, we present these works sorted by dialect work frequency:

- **Egyptian dialect**

In 2012, Habash N. et al. developed **CALIMA** (Habash, Eskander, & Hawwari, A morphological analyzer for Egyptian Arabic, 2012) a tool for morphological analysis of the Egyptian dialect which relies on ECAL (Kilany, Gadalla, Arram, Yacoub, & ElHabashi, 2002) (an Egyptian dialect lexicon). It follows the POS guidelines used by the linguistic Data Consortium for Egyptian (Maamouri, Krouna, & Tabessi, 2012) and accepts a variety of orthographic spelling normalized to the conventional orthography of Arabic dialects CODA (Habash, Diab, & Rabmow, Conventional Orthography for Dialectal Arabic, 2012). CALIMA has 100K stems corresponding to 36K lemmas. Evaluation of this analyzer was performed using 3300 manually annotated words of an Egyptian corpus and showed a correct answer for POS tags over 84% of the time.

In a later work (2013), Habash N. et al. built **MADA-ARZ** (Habash, Roth, Rambow, & Eskander, 2013) the Egyptian Arabic morphological analyzer. They used the MSA version MADA (Habash, Rambow, & Roth, MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization, 2009) and retrained the EGY annotations of $\text{CALIMA}_{\text{EGY}}$ analyzer. They showed that it is useful in building machine translation from Egyptian dialect to English. This analyzer can be considered as a new version of MADA developed specifically for the Egyptian dialect.

In 2014 Shalloum W. and Habash N. built **ADAM** (Salloum & Habash, ADAM: Analyzer for Dialectal Arabic Morphology, 2014) on the top of SAMA (Graff, et al., 2009) database. It can analyze both of Egyptian and Levantine dialects. In their adopted approach, they extended SAMA database through adding dialectal affixes and clitics, in addition to a set of handwritten rules. ADAM follows the same database format used in the MSA analyzer ALMOR (Habash, Arabic Computational Morphology. Knowledge-based and Empirical Methods, 2007) and outputs analyzed text as lemma and feature-value pairs including clitics. This analyzer is intended for improving machine translation performance. As an example, it is used as part of ELISSA which translates dialectal text to MSA. The experimental results show that the out of vocabulary rates to 16,1% in Levantine and 33,4 for Egyptian texts.

After that in 2014, Arfath Pasha et al. developed a new version of MADA, called **MADAMIRA** (Pasha, et al., 2014) for both MSA and EGY. They added also AMIRA (Diab & Jurafsky, 2007) features to this new analyzer. After cleaning and preprocessing input text, the system uses feature modeling component in order to apply SVM and language models which helps to derive predictions for the word's morphological feature. One important benefit of MADAMIRA is the analysis ranking component which scores each word's analysis list according to model predictions, and then sorts the analyses as output text based on scores. MADAMIRA can easily be integrated in web applications. Accuracy reached in evaluation exceeds 83%.

Finally, authors of (Khalifa, Zalmout, & Habash, 2016) combined two MSA Morphological analyzers: MADAMIRA and FARASA (Darwish & Mubarak, 2016) to build YAMAMA morphological analyzer for Egyptian dialect. In fact, they used from MADAMIRA a component analyzing text without context reading, and inspired from FARASA the disambiguation modeling component. YAMAMA produce the same output as MADAMIRA and reached 79% of accuracy. Despite its low accuracy, it is five time faster compared to MADAMIRA.

- **Levantine dialect**

In addition to ADAM described above, Levantine dialect was targeted by MAGEAD analyzer (Habash & Rambow, 2006). In 2006, Habash N. and Rambow O. focused their efforts on modeling Arabic dialects directly by building MAGEAD analyzer which can decompose word forms into the templatic morphemes and relates morphemes to string. The principle of its analyses relies on lexeme and features. They define the lexeme as a triple containing a root, a meaning index and a morphological behavior class (MBC). The first version of this dialect analyzer covers Levantine dialect in addition to MSA.

Similarly, Eskandar R. et al. (Eskander, Habash, Rambow, & Pasha, 2016) started from annotated corpora and MADAMIRA models to build two morphological analyzers: one for EGY ($\text{ALMOR}_{\text{EGY}}$) and the other for LEV ($\text{ALMOR}_{\text{LEV}}$). They used the Egyptian Arabic corpora (Maamouri, Krouna, & Tabessi, 2012) as the EGY data and Curras Corpus of Plestinian Arabic (Jarrar M. , Habash, Alrimawi, Akra, & Zalmout, 2016) as he LEV data. These corpora are morphologically annotated in similar style to the

annotations in MADAMIRA database (ALMOR). The morphological analyzers were created for different sizes from 5K up to 135K where the big analyzers sizes are of 135K for EGY and of 45K for LEV. When evaluating the analyzers, $\text{ALMOR}_{\text{EGY}}$ reached an accuracy of 90% while $\text{ALMOR}_{\text{LEV}}$ reached only 87%.

TABLE I

ADVANTAGES AND DISADVANTAGES OF SURVEYED AD MA

| Morphological Analyzer | Advantages | Disadvantages |
|---|---|---|
| $MAGEAD_{LEV}$ | rich linguistic information and rules | low accuracy.<br>time consuming to extend. |
| $CALIMA_{EGY}$ | Extensible to new dialects | low coverage of dialect words. |
| $AlKhalil_{GLF}$ | multi-dialects + extensible to new dialects | Uses commercial components.<br>text context is missed. |
| $ALMOR_{YEM}$ | Extensible to new dialects | time consuming.<br>text context is missed. |
| $BAMA_{ALG}$ | rich linguistic information | time consuming.<br>acceptable accuracy.<br>text context is missed. |
| $MAGEAD_{TUN}$ | High lexical coverage | misses syntactic component.<br>text context is missed |
| MADA-ARZ | includes a ranking score analysis according to context. | It needs advanced skills to explore its functions. |
| ADAM | Extensible to other dialects | low accuracy<br>text context is missed |
| $MADAMIRA_{EGY}$ | includes a ranking score analysis according to context.<br>extensible to new dialects | to extend MADAMIRA, annotations need to be added manually + may not analyze some cases + slow processing |
| $AlKhalil_{TUN}$ | rich linguistic information | time consuming<br>text context is missed |
| YAMAMA | Fast processing | low accuracy |
| $CALIMA_{GLF}$ | rich linguistic information | Limited to verbs only.<br>text context is missed |
| $ALMOR_{LEV}$ | includes a ranking score analysis according to context.<br>extensible to new dialects | Limited database |
| $ALMOR_{EGY}$ | | |

- **Gulf dialects**

Almeman K. and Lee M. (Almeemam & Lee, 2012) used Alkhalil in 2012 by adding dialect affixes to its database. They splitted processing on two steps. In the first one, Alkhalil was able to analyze dialect words sharing the same stem with MSA words with an accuracy of 69%. If the system can't produce corresponding analysis with existing database, then it segments input text and uses 'the web as corpus' to estimate frequency of different segment combinations. These items were used to guess the correct base form. The overall synthesis is shown to have 69% accuracy on a corpus of Gulf dialects.

Khalifa S. et al. (Khalifa, Hassan, & Habash, 2017) developed $CALIMA_{GLF}$ a morphological analyzer for Emirati (EMR) Arabic verbs. They used two resources that provide explicit linguistic knowledge. The first is a database gathering a collection of roots, patterns and affixes. While the second consists of a lexicon specifying verbal entries with roots and patterns. By merging these two resources in one model, all possible analyses are provided to cover over 2600 EMR verbs and following MADAMIRA and $CALIMA_{EGY}$ representation. Evaluation result of $CALIMA_{GLF}$ gives 81% of accuracy.

**Yemeni dialect**

Al-Shargi F. (Al-Shargi, Kaplan, Eskander, Habash, & Rambow, May 2016) performed in 2016 an effort to annotate dialect corpora in order to adapt MADAMIRA to Yemeni dialect. They used DIWAN (Rambow & Al-Shargi, 2015) to manually annotate corpora collected from both online and printed materials which rated to 32.5K words. This annotated corpus was used then to adapt ALMOR MSA to the Yemeni dialect by extending its database to cover this dialect. The overall evaluation of the new analyzer built rated to 69.3%. $MAGEAD_{LEV}$

- **Tunisian Dialect**

Zribi I. et al. (Zribi, Ellouze Khemakhem, & Hadrich Belguith, 2013) adapted in 2013 Alkhalil morphological analyzer (Boudlal A., 2010) to Tunisian dialect. They first built a corpus and a lexicon by recording speech and manually transcribing some radio and TV broadcasts. They integrated the Tunsian lexicon in Alkhalil process and then added Tunisian linguistic rules such as roots and patterns. This task was time consuming given that each root must be combined with corresponding patterns. The system performance resulted on the built corpus in the first step rates to 77%.

In another version of MAGEAD (2014), Hamdi A. et. al (Hamdi, Núria, Alexis, & Habash, 2014) extended this analyzer to cover Tunisian dialect by converting dialectal text to a pseudo MSA form to achieve 82% of accuracy.

- **Algerian dialect**

Harrat S. et al. (Harrat, Meftouh, Abbas, Hidouci,, & Smaili, 2016) extended in 2016 BAMA (Buckwalter, 2002) analyzer to Algerian dialect using an Algerian lexicon and added necessary affixes and stems to BAMA database. The new analyzer reached an accuracy of 69% when evaluating it on an Algerian corpus.

- **Summary**

Finally, we summarize in Table I the advantages and disadvantages of each morphological analyzer surveyed above. In the next section we present approaches that have been followed to build them.

## IV. AD Morphological Analysis Techniques

From our previous reviews, we can consider that works performed in order to provide morphological analyzers for Arabic dialects generally fall in two camps. The first trend gathers solutions modeling Arabic dialects directly. They are built from scratch and contain rich linguistic representations and morphological rules. Moreover, these systems rely on lexicons and compile the effect of the morphemic, phonological and orthographic rules in this lexicon itself. As an illustration, MAGEAD and CALIMA approaches follow this direction. The benefit of applying this approach is the valuable performance

| | | |
|---|---|---|
| EGY sentences | مالك ومال اصحاب المعالى يا اخى الفاضل زعلان ليه | 1 |
| | لانك شخصيه وبجد مش هعرف اوصفها | 2 |
| | المفروض ان اي اتنين بيحبوا بعض ميجرحوش بعض | 3 |
| TUN sentences | شبيك و شمدخلك في اصحاب المعالي يا خويا لعزيز علاش متغشش | 1 |
| | على خاطرك شخصية بلحق منجمش نوصفها | 2 |
| | المفروض اي زوز يحبو بعضهم ما يجرحوش بعضهم | 3 |
| MSA translation | ما مشكلتك مع سيادته يا أخى العزيز لما انت منزعج هكذا | 1 |
| | لأنك شخصية و لن اقدر أن أوصفها | 2 |
| | و من المفترض ان اى شخصين فى علاقه حب لا يجرحوا بعضهما البعض | 3 |
| English translation | What you have with his Excellency, my dear brother, why are you upset (sad) for? | 1 |
| | because you are a personality that i can not describe | 2 |
| | Its supposed to be that any two that loves each other not hurt each other | 3 |

Fig.3 Sample of multi-Arabic dialect corpus

resulted on testing these systems. However, their main drawback is the important efforts needed in their building and extending for other Arabic dialects (time consuming).

The second direction followed to build Arabic dialect MA includes systems extended from existing MSA analyzers. It consists of either superficial or deep adaptation of existing MSA analyzers. The first is a light modification usually related to database or the included lexicon in the MSA analyzer. For example, in the work of K. Almeemam (Almeemam & Lee, 2012), only affixes table was concerned by the task of Alkhalil adaptation to Arabic dialects. While the second relies on a deep process concerns in addition to the AD MA database, the language modeling and the used algorithm in the MSA analyzer. As an illustration, the work of MADAMIRA falls in this category.

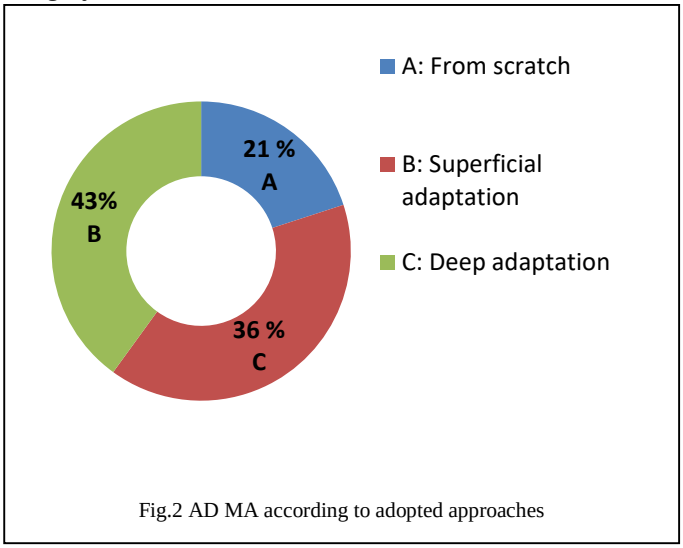


Fig.2 AD MA according to adopted approaches

We provide in Figure 2 and Table II the surveyed AD morphological analyzers categorized according to the adopted approach.

If we consider the fact that deep and superficial adaptation uses the same approach's concept (adaptation of MSA analyzers to AD), we can affirm the dominance (79%= 36% + 43%) of this approach over the other technique (building from scratch) as shown in figure 2. Hence, it seems that researchers prefer to adapt existing MSA analyzers to Arabic dialects due to time consuming.

TABLE II
ARABIC DIALECT MORPHOLOGICAL ANALYZERS BY APPROACHES

| From scratch approach | Adaptation approach | |
|---|---|---|
| | Superficial | Deep |
| $MAGEAD_{LEV}$ | $AlKhalil_{GLF}$ | MADA-ARZ |
| $CALIMA_{EGY}$ | $ALMOR_{YEM}$ | ADAM |
| $CALIMA_{GLF}$ | $BAMA_{ALG}$ | MADAMIRA |
| | $MAGEAD_{TUN}$ | $AlKhalil_{TUN}$ |
| | $ALMOR_{LEV}$ | $YAMAMA_{EGY}$ |
| | $ALMOR_{EGY}$ | |

## V. Benchmark of AD MA

To compare AD MA performances, it is necessary fist that the AD MA being compared be freely available in addition to preparing a multi AD evaluation corpus. Unfortunately, not all AD MA surveyed in section 3 are available nor the evaluation corpus. For this reason, we decided on one hand to limit the scope of this benchmark to only available AD MA that are $MADAMIRA_{EGY}$, $ALKHALIL_{TUN}$ and YAMAMA. On other hand, we decided to select and annotate a portion of an existing multi AD corpus (Bouamor, Habash, & Oflazer, 2014). This benchmark follows the same general design as the work of (Jaafar, Bouzoubaa, Yousfi, Tajmout, & Khamar, 2016) where the considered MA are designed for MSA. The following steps are carried out in order to ensure our benchmark:

1) Getting AD MA resources: $MADAMIRA_{EGY}$ was downloaded from official web site, where $ALKHALIL_{TUN}$ and YAMAMA were obtained by contacting their authors. Hence, this benchmark considers only Egyptian and Tunisian dialects.

2) Preparing an evaluation corpus: From the multi AD corpus cited above, we selected 10 sentences corresponding to the two studied dialects (EGY and TUN). Then, we annotated each word using first an automatic processing then a manual validation. The first one is performed using corresponding AD MA that generates possible analyses of each word. The second one is ensured manually where we correct manually the previous results. Figure 3 gives an example about the annotation of the same sentence in each dialect.

3) Experiments and metrics set up: After obtaining the three AD MA and preparing the evaluation corpus, we ran selected

sentences on these analyzers and generated each analyzer results. Thus, each word in the evaluation corpus has several analyses in each analyzer.

In order to evaluate these results, we used common evaluation metrics namely: precision, recall, accuracy and F-measure. However, these metrics do not consider time taken in order to process input text (run time). For this reason, we used the $G_M score$ introduced in *(Jaafar, Bouzoubaa, Yousfi, Tajmout, & Khamar, 2016)* which gathers morphological tags, accuracy and run time to evaluate given morphological analyzer. This score is obtained by applying the following formula:

$$G_M\text{score} = \frac{\text{RT}}{AC + STg + \alpha.ATg}$$

Where:

- RT: run time.
- AC: accuracy of AD MA results.
- STg: count of standard tags considered by the AD MA which are: vowelized form, stem, pattern, root, POS, prefix and suffix.
- ATg: count of additional tags considered by the AD MA.

In our experiment, we set α=1/4 in order to decrease the weight of additional tags in the $G_M$score because we consider that standard tags are more important compared to additional tags. Note that, among compared AD MA, the one having the lowest $G_M$score is considered as the best. Moreover, when $G_M$score value tends to zero the AD MA is perfect. Table III presents, for each compared AD MA, obtained results related to the metrics described above.

TABLE III
ARABIC DIALECT MORPHOLOGICAL ANALYZERS BY APPROACHES

| Metrics | MADAMIRA | ALKHALIL | YAMAMA |
|---|---|---|---|
| # analyzed words | 105 | 105 | 105 |
| # words not analyzed | 1 | 21 | 1 |
| Run time | 23 | 28 | 12 |
| Precision | 0,86 | 0,68 | 0,86 |
| Recall | 0,88 | 0,66 | 0,88 |
| Accuracy | 0,85 | 0,68 | 0,86 |
| F-measure | 0,87 | 0,67 | 0,87 |
| Standard Tags | 5 | 7 | 5 |
| Additional Tags | 12 | 7 | 12 |
| $G_M$score | 2,59 | 2,96 | 1,35 |

By looking at the benchmark results presented in Table III, it seems that YAMAMA analyzer gives the best results since it has the lowest $G_M$score. This analyzer holds the top position thanks to its lowest run time followed by MADAMIRA then $\text{AlKhalil}_{\text{TUN}}$. Regarding covered tags, $\text{AlKhalil}_{\text{TUN}}$ returns all standard tags and five additional tags, where MADAMIRA and YAMAMA return only five required tags but return also 12 additional tags. If we consider the accuracy, MADAMIRA and YAMAMA are more accurate than $\text{AlKhalil}_{\text{TUN}}$ and give approximatively the same accuracy.

TABLE IV
ARABIC DIALECT MORPHOLOGICAL ANALYZERS (TIMELINE AND ACCURACIES)

| Arabic Dialect | Morphological Analyzer | Accuracy | Year |
|---|---|---|---|
| **EGY** | CALIMA | 84% | 2012 |
| | MADA | - | 2013 |
| | ADAM | 67% | 2014 |
| | MADAMIRA | 83% | 2014 |
| | YAMAMA | 79% | 2016 |
| | ALMOR | 90% | 2016 |
| **LEV** | MAGEAD | 56% | 2006 |
| | ADAM | 84% | 2014 |
| | ALMOR | 87% | 2016 |
| **TUN** | AlKhalil | 77% | 2013 |
| | MAGEAD | 82% | 2014 |
| **YEM** | ALMOR | 69% | 2016 |
| **GLF** | AlKhalil | 69% | 2012 |
| **ALG** | BAMA | 69% | 2016 |
| **EMR** | CALIMA | 81% | 2017 |

## VI. DISCUSSION

Table IV summarizes the related research work on building Arabic Dialect analyzers. It presents general informations about different AD MA surveyed in section III. Based on this table and on the descriptions in the previous sections the following comments can be made.

- **Performance difference**: AD Morphological Analyzers reached low accuracy compared to MSA even if AD ones are extended from MSA analyzers using the same components (except lexicons). This fact can be observed looking at the benchmark results in the work of (Jaafar, Bouzoubaa, Yousfi, Tajmout, & Khamar, 2016) where the MSA morphological analyzers have a $G_M$score less than 1 but in our benchmark the lowest value of AD MA $G_M$score rates to 1,35. Knowing that these AD MA are extended on the basis of the MSA analyzers.
  This result can be explained by the fact that MSA presents high level of standardization, syntactic and grammatical rules. Whereas, Arabic dialects still integrate new lexical lemma and grammatical rules especially from foreign languages.
- **AD Coverage**: existing AD analyzers do not cover all Arabic dialects. In addition, some covered dialects are strongly addressed compared to other dialects such as Egyptian and Levantine dialects. This may be explained by the fact that Egyptian and Levantine dialects are the most popular Arabic dialects since they are the most used in Arabic media. As a matter of fact, first Arabic dialect lexicons made were addressed to these dialects which was helpful to deal with resources lack and build corresponding analyzers. Nevertheless, some other AD lexicons increasingly are available (Shoufan & Alameri, 2015). This

can be useful in the future to address remaining Arabic dialects in morphological analysis.

- **Approaches**: researchers usually prefer to adapt existing MSA analyzers (79%) than build AD MA from scratch (21%) which reflect relatively the similarities between MSA and AD. One reason to explain this orientation is to avoid time consuming if they follow the second option. In fact, they benefit from the closeness existing between MSA and Arabic dialects especially in lexical level in order to extend MSA analyzers to AD. As an example, 81% of Moroccan dialect lexicon is borrowed from Arabic according to authors of (Tachicart, Bouzoubaa, & Jaafar, Lexical differences and similarities between Moroccan dialect and Arabic, 2016). Hence, it seems suitable adapting MSA tools for this dialect rather than starting from scratch. Note that obtained results in the process of adaptation are encouraging even if resulted analyzers output with lower quality.
- **NLP integration**: Using these AD analyzers in large scale NLP systems is not reached yet. Currently, they are only integrated in machine translation or automatic identification systems. In fact, processing Arabic dialects is still in earlier stages compared to MSA. Moreover, researchers focalize their effort in building resources and basic NLP systems such as morphological analyzers and machine translation till now. In the future, they will give more attention to advanced NLP systems thanks to the availability of resources and MA which are important to this end.
- **Adaptation of MSA analyzers**: In the following we describe some directions to adapt most important morphological analyzers to new Arabic dialects:
  - **MADAMIRA**: Currently, this analyzer can process both MSA and EGY. In order to extend it to new AD, researchers need to integrate corresponding dialect lexicon following MADAMIRA database format. In addition, it is necessary to create their own language model and statistical classifiers in order to replace MSA components.
  - **AlKhalil**: this analyzer does not include any lexicon. However, it is necessary to integrate necessary dataset composed of roots and patterns of the dialect to be processed. Then, it requires to express all possible combinations between them in order to represent all possible lemma. Note that the adaptation process of this analyzer requires an important effort compared to MADAMIRA.

## VII. Conclusion

In this paper we have surveyed Arabic dialect morphological analyzers. We have presented their advantages and disadvantages with detailed description. We described also followed approaches to build such analyzers, and then we classified surveyed AD MA according to these approaches. In the performed benchmark on AD MA, we found that MSA analyzers work better than Arabic dialect ones instead of the majority of AD MA are extended from MSA ones. We observed also a difference between AD MA reached accuracies. We believe that Arabic dialects increasingly keep more attention which can in the future improve the performance and coverage of Arabic dialect morphological analyzers.

## Acknowledgment

We would like to thank Nizar Habash, Ines Zribi and Sallam Khalifa for sharing with us their data including morphological analyzers which was necessary to perform the benchmark process.

## References

Almeemam, K., & Lee, M. (2012). Towards Developing a Multi-Dialect Morphological Analyser for Arabic. *4th International Conference on Arabaic Language Processing (CITALA'12).* Rabat.

Al-Shargi, F., Kaplan, A., Eskander, E., Habash, N., & Rambow, O. (May 2016). Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic. *10th Language Resources and Evaluation Conference (LREC 2016).* Portoroz, Slovenia.

Authore, F. (5-6 November 2017). From Algerian Dialect into MSA: Neural and Statistical based Machine Translation. *3rd International Conference on Arabic Computational Linguistics, ACLing 2017.* Dubai, United Arab Emirates.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Jonathan Wright, H. L., Strassel, S., . . . Rambow, O. (2014). Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. *EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, (pp. 93–103). Doha, Qatar.

Bouamor, H., Habash, N., & Oflazer, K. (2014). A Multidialectal Parallel Corpus of Arabic. *Ninth International Conference on Language Resources and Evaluation (LREC-2014).* Reykjavik, Iceland.

Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., & Lakhouaja, A. (2014). L'Analyseur Morphosyntaxique AlKhalil Morpho Sys 2. *1ère Journée Doctorale Nationale sur L'Ingénierie de la Langue Arabe, (JDILA'14), Rabat, Maroc, 2014.* Rabat.

Boudlal A., L. A. (2010, December). AlkhalilMorpho Sys1: A Morphosyntactic analysis system for Arabic texts.

Buckwalter, T. (2002). Buckwalter arabic morphological analyzer version 1.0.

Darwish, K., & Mubarak, H. (2016). Farasa: A New Fast and Accurate Arabic Word Segmenter. *Tenth International Conference on Language Resources and Evaluation.* Portorož, Slovenia.

Diab, M. H., & Jurafsky, D. (2007). Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking. *Arabic Computational Morphology: Knowledge-based and Empirical Methods.Kluwer/Springer Publications.*

Elfardy, H., Al-Badrashiny, M., & Diab, M. (2014). AIDA: Identifying code switching in informal Arabic text. *EMNLP 2014: Conference on Empirical Methods in Natural Language Processing* , (p. 94). Doha.

Elfardy, H., Al-Badrashiny, M., & Diab, M. (2014). AIDA: Identifying Code Switching in Informal Arabic Text. *1st Workshop on Computational Approaches to Code Switching* (pp. 94-101). Doha, Qatar: Association for Computational Linguistics.

Eskander, R., Habash, N., Rambow, O., & Pasha, A. (2016). Creating Resources for Dialectal Arabic from a Single Annotation: A Case Study on Egyptian and Levantine. *COLING 2016, 26th International Conference on Computational Linguistics*, (pp. 3455-3465). Osaka, Japan.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., & Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1.

Habash, N. (2007). *Arabic Computational Morphology. Knowledge-based and Empirical Methods.* (A. v. Soudi, Ed.) Kluwer/Springer.

Habash, N., & Rambow, O. (2006). Magead: A morphological analyzer and generator for the arabic dialects. *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistic*, 681–688.

Habash, N., Diab, M., & Rabmow, O. (2012). Conventional Orthography for Dialectal Arabic. *Language Resources and Evaluation Conference (LREC).* Istanbul.

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological analyzer for Egyptian Arabic. *Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, (pp. 1-9).

Habash, N., Rambow, O., & Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. *the Second International on Arabic Language Resources and Tools.* The MEDAR Consortium.

Habash, N., Roth, R., Rambow, O., & Eskander, R. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).* Atlanta, GA.

Hamdi, A., Núria, G., Alexis, N., & Habash, N. (2014). POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools. *ProceedingsoftheSecondWorkshoponArabicNaturalLanguageProcessing*, (pp. 59 - 68). Bejing.

Harrat, S., Meftouh, K., Abbas, M., H. K., & Smaili, K. (2016). An Algerian dialect: Study and Resources. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *7*(3), 384 - 395.

Jaafar, Y., Bouzoubaa, K., Yousfi, A., Tajmout, R., & Khamar, H. (2016). Improving Arabic morphological analyzers benchmark. *International Journal of Speech Technology*, *19*(2), 259–267.

Jarrar, M., Habash, N., Akra, D., & Zalmout, N. (2014). Building A Corpus For Palestinian Arabic: A Preliminary Study. *EMNLP 2014 Workshop on Arabic Natural Language Processing. Association for Computational Linguistics (ACL)*, (pp. 18-27). Doha, Qatar.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., & Zalmout, N. (2016, December). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, *I*(10579), 1-31.

Khalifa, S., Hassan, S., & Habash, N. (2017). A Morphological Analyzer for Gulf Arabic Verbs. *The Third Arabic Natural Language Processing Workshop.* Valencia, Spain.

Khalifa, S., Zalmout, N., & Habash, N. (2016). YAMAMA: Yet Another Multi-Dialect Arabic Morphological Analyzer. *the 26th International Conference on Computational Linguistics.* Osaka.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., & ElHabashi, A. (2002). Egyptian Colloqium Arabic Lexicon.

Maamouri, M., Bies, A., Kulick, S., Krouna, S., Tabassi, D., & Ciul, M. (2012). Egyptian Arabic Treebank DF Parts 1-8.

Maamouri, M., Krouna, S., & Tabessi, D. (2012). *Egyptian Arabic Morphological Annotation Guidelines.*

May, J., Benjira, Y., & Echihabi, A. (2014). An Arabizi-English Social Media Statistical Machine Translation System Machine Translation in the Americas. *the Eleventh Biennial Conference of the Association for* . Vancouver, Canada.

Mona Diab, N. H. (2010). COLABA: Arabic dialect annotation and processing. *LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*, 66-74.

Pasha, A., Al-Badrashiny, M., ElKholy, A., Eskander, R., Diab, M., Habash, N., . . . Roth, R. (2014, May). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *the Ninth International Conference on Language Resources and Evaluation.* Reykjavik.

Rambow, O., & Al-Shargi, F. (2015, July). DIWAN: A Dialectal Word Annotation Tool for Arabic. *The Second Workshopon Arabic Natural Language Processing.* Beijing.

Salloum, W., & Habash, N. (2012). Elissa: A Dialectal to Standard Arabic Machine Translation System. *COLING: International Conference on Computational Linguistics*, (pp. 385-392). Mumbai.

Salloum, W., & Habash, N. (2014, December). ADAM: Analyzer for Dialectal Arabic Morphology. *Journal of King Saud University - Computer and Information Sciences.*, *Volume 26*(Issue 4), Pages 372–378.

Shoufan, A., & Alameri, S. (2015). Natural Language Processing for Dialectical Arabic: A Survey. *the Second Workshop on Arabic Natural Language Processing.* Beijing.

Tachicart, R., Bouzoubaa, K., & Jaafar, H. (2014, November). Building a Moroccan dialect electronic Dictionnary

(MDED). *5th International Conference on Arabic Language Processing CITALA.* Oujda.

Tachicart, R., Bouzoubaa, K., & Jaafar, H. (2016). Lexical differences and similarities between Moroccan dialect and Arabic. *4th IEEE International Colloquium on Information Science and Technology (CiSt).* Tanger.

Tachicart, R., Bouzoubaa, K., Aouragh, S. L., & Jaafar, H. (2017). Automatic Identification of the Moroccan Colloquial Arabic. *6th International Conference on Arabic Language Processing ICALP 2017.* Fez, Morocco: Springer.

Zribi, I., Ellouze Khemakhem, M., & Hadrich Belguith, L. (2013). Morphological Analysis of Tunisian Dialect. *International Joint Conferenceon Natural Language Processing*, (pp. 992–996). Nagoya.