Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Moroccan Arabic vocabulary generation using a rule-based approach

Ridouane Tachicart *, Karim Bouzoubaa

Mohammadia School of Engineers, Mohammed V University in Rabat, Avenue Ibn Sina B.P 765 Agdal Rabat, 10090, Morocco

ARTICLE INFO

Article history:
Received 20 August 2020
Revised 16 February 2021
Accepted 26 February 2021
Available online xxxx

Keywords:
Generation
Morphology
Moroccan Arabic
Corpus
Lexicon
Natural language processing
Morphological analyzer
Standard Arabic

ABSTRACT

NLP resources play a crucial role in the building of many NLP applications. The importance of these resources depends not only on their size and coverage but also on the richness and the precision of the annotated information they provide. In the case of resource-scarce languages such as Moroccan Arabic, the building of NLP applications is limited due to the lack of these resources. To overcome this problem, we follow a rule-based approach to generate a Moroccan morphological vocabulary (MORV) which constitutes the first step addressing the problem of Moroccan morphological generation. MORV is designed and implemented based on two main components: On one hand, an MA lexicon and a list of fully annotated affixes and clitics that we have created specifically to ensure the generation process. On the other hand, a set of rules covering the concatenation and the orthographic adjustments of the generated words. Moreover, given a base form, MORV outputs more than 4.5 M Moroccan words with rich morphological features such as tense, gender, number, state, etc. We tested the coverage of MORV on texts collected from Moroccan social media and realized that it reaches a vocabulary coverage of 84% and a precision of 94%. This system is a benefit for building other NLP applications such as spell checking, morphological analysis, and machine translation.

1. Introduction

Morphology is the field of linguistics that studies the internal structure of words. The main purpose of morphology is to analyze the word structure and to describe the meaningful units called morphemes (atomic linguistic units that carry meaning) for a given word. From a practical point of view, the word structure can be expressed using different morphological attributes such as tense, person, number, gender, etc.

One important implementation of morphology is vocabulary generation. It is the process of word-formation that produces inflected forms of a word starting from labelled morphemes. Regarding the Arabic language that includes Modern Standard Arabic (MSA) and a set of dialects, there are two main approaches to the morphological generation of vocabularies (Habash, 2010): either following concatenative or templatic morphology. In order to create a word, concatenative morphology involves the combination of lemma or stem with morphemes such as affixes and clitics, whereas templatic morphology is interleaving and merges roots with patterns.

Lexical resources are crucial and important to most NLP applications such as morphological analyzers. However, the Moroccan Arabic dialect (MA) is considered a resource-scarce language since it suffers from the lack of available MA resources and NLP tools. In fact, there are currently no morphological analyzers nor morphological generation systems. Thus, analyzing MA texts is restricted to manual tasks. Moreover, various NLP applications rely on extracting the morphological information encoded in the word. Additionally, if we consider that some NLP applications suffer from data sparsity such as statistical machine translation, the availability of MA resources exhibiting morphologically annotated MA words can alleviate these problems. Hence, building a new resource describing MA words morphology is useful to facilitate the building of NLP applications such as morphological analyzer and machine translation.

Previous approaches to generate vocabularies for standard languages followed either statistical (Faruqui et al., 2016) & (Dusek and Jurcícek, 2013) or rule-based techniques (Bauer, et al., 2015) & (Viks, 2000) & (Jisha, et al., 2011). The first relies on training taggers on large annotated corpus using common machine learning algorithms such as Support Vector Machines (SVM) (Vapnik,

* Corresponding author.
E-mail addresses: ridouane.tachicart@research.emi.ac.ma (R. Tachicart), bouzoubaa@emi.ac.ma (K. Bouzoubaa).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.02.013
1319-1578/© 2021 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

Please cite this article as: R. Tachicart and K. Bouzoubaa, Moroccan Arabic vocabulary generation using a rule-based approach, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2021.02.013

1995) or LSTM (Hochreiter and Schmidhuber, 1997), etc.). Saving time and increasing accuracy are the main advantages of this approach. Unfortunately, no such resources are currently available to train MA taggers. The second is rule-based and consists in using lexicons of morphemes and implementing decision algorithms, using for example finite-state transducers (FSTs). The latter govern the concatenation of different morphemes and output new words with their morphological analysis. Such an approach is appealing since it meets the linguistic requirements. Furthermore, with the lack of currently annotated MA corpora, following this approach seems to be the most suitable solution to generate MA morphological vocabulary.

The main contribution of this paper is to present and evaluate our MORphological Vocabulary (MORV) using a MORphological Generator (MORG) that relies on a rule-based approach. The idea behind conducting such a work is to establish a Moroccan Arabic morphological analyzer that will enable solving various NLP tasks.

In our method, we used a lexicon of MA lemmas and a table of MA annotated morphemes (affixes and clitics) as a dataset. Besides, we stored linguistic and orthographic rules in separate tables to seamlessly govern the concatenation of different morphemes by an appropriate algorithm. MORV evaluation consists of assessing the generated output regarding two aspects. The first (quantitative evaluation) aims at ensuring that MORV entries (generated words only) cover sufficiently the Moroccan Arabic dialect. While the second involves assessing the precision of the morphological information that MORV provides using common evaluation metrics such as Precision, Recall, and F-measure. In this perspective, the main advantage of MORV is the good coverage of the Moroccan Arabic vocabulary, the flexibility in managing rules and the ability to be easily extended.

In section 2 we discuss related works dealing with vocabulary generation. In section 3, we exhibit morphological information about different MA categories. After that in section 4, we highlight two linguistic approaches to Moroccan Arabic word generation. In section 5, we discuss the main objectives of building MORV. Then, we present the adopted approach and its implementation. We present the result of MORV evaluation and discuss its features in section 6. Finally, we conclude this paper with some perspectives in section 7.

2. Related work

Unlike the Arabic language, a few works are dealing with the morphological generation of Arabic dialects vocabularies. In addition, there is currently no work addressing the morphological generation of MA vocabularies to the best of our knowledge. The literature review of Arabic morphological vocabularies exhibits various approaches that can be mainly classified into manual annotations such as in (Al-Shargi et al., 2016), (Maamouri et al., 2006) and automatic approaches. In the following, we summarize related works concerning the automatic generation of both MSA and dialectal morphological vocabularies.

Among the earliest efforts to build Arabic morphological generators was the work of Beesley (Beesley, 1996) & (Beesley, 2001) using Xerox's finite-state transducer[1]. To implement the generator, the author compiled a lexical database including 4930 roots and 400 patterns as well as a set of morphotactics and alternations rules that govern the combination of stems with clitics. The result of running the system over the lexical database gives 72 M fully inflected forms.

In the work of Cavalli-Sforza et.al (Cavalli-Sforza, et al., 2000) which is reviewed also in (Soudi, et al., 2007), authors presented an approach to generate Arabic verbs using MORPHE (Leavitt, 1992), a tool for modeling morphology based on discrimination trees and regular expressions. The system follows the concatenative morphology and is driven by a morphological form hierarchy governing not only the relationship between roots and patterns forms but also transformational rules that attach to leaf nodes in the hierarchy.

Habash (Habash, 2004) presented Aragen as a lexeme-based Arabic morphological generator that follows concatenative morphology. Aragen uses Buclwalter's database (BAMA) (Buckwalter, 2002) that includes a set of tables representing morphotactics and orthographic rules. In this database, we find a lexicon of annotated morphemes (lemma, affixes and clitics) and a compatibility morphemes table that indicates which morpheme can be concatenated to which other. To evaluate Aragen, the author used a sample of 1 M words from the UN Arabic-English corpus (Jinxi, 2002) and realized that it reaches a coverage of 76%.

Authors in (Habash et al., 2005), (Habash & Rambow, 2006) and (Habash & Rambow, 2007) built MAGEAD a morphological generator and analyzer of Modern Standard Arabic (MSA) and Levantine (LEV) verbs using FSTs. MAGEAD follows templatic morphology where the principle of its analysis relies on lexeme and features. Authors define the lexeme as a triple containing a root, a meaning index and a morphological behavior class (MBC). In another work (Altantawy, et al., 2010), MAGEAD has been extended to cover MSA nouns and adjectives. It should be noted that MAGEAD is the first tool for Arabic dialects that includes roots and patterns in its implementation. Also, it was very helpful in the process of corpora annotation in several works such as the work of (Diab, et al., 2010).

Shaalan, et al. (Shaalan et al., 2007) performed an effort to build a rule-based Arabic morphological generator in order to facilitate the task of automatic translation. Using the logic programming language Prolog, authors implemented this generator by encoding transformational rules that govern the concatenation of affixes with Arabic lemmas.

(Attia, et al., 2011) & (Attia, et al., 2014) developed AraComLex an MSA morphological processing toolkit based on finite-state transducers. The implementation of AraComLex follows the concatenative morphology and considers the lemma as the base form. The authors used a lexical database containing more than 30 k lemmas in order to generate about 9 M surface forms. In another work, Shaalan et. al (Shaalan et al., 2012) created an open-source resource of Arabic words on the basis of AraComLex transducer. The main goal of building this resource is to facilitate the building of an Arabic spelling checker. Authors used Microsoft spell-checker (included in Microsoft Office 2010) to validate a set of 9 M words from 13 M AraComLex generated words.

Neme (Amid Neme, 2013) built a vocabulary of 2.5 M Arabic verbs starting from 15.4 K verbs and following templatic morphology by using finite-state transducers (FSTs). To evaluate the generated vocabulary, the author used 10 K verbs extracted from NEMLAR corpus (Attia, et al., 2005). It should be noted that the accuracy rate is not reported.

Doumi et. al (Doumi, et al., 2016) built a lexical resource that contains 11 M verbal inflected forms. They followed a concatenative morphology and used for this purpose a representative corpus of MSA to extract verb lemmas. They used a corpus instead of a lexicon in order to avoid obsolete words that have no place in current usage. Then, they used FSTs in order to generate MSA verbs following MSA concatenation rules with orthographic adjustments. Evaluation results showed that the generated resource covers more than 70% of the MSA verbs.

Khalifa et. al (Khalifa, et al., 2017) introduced CALIMA$_{GLF}$ as a morphological analyzer and generator for Emirati (EMR) Arabic verbs. In this work, two resources providing explicit linguistic knowledge are used. The first is a database gathering a collection

[1] https://web.stanford.edu/~laurik/fsmbook/home.html

of root-abstracted paradigms that map from features to root-abstracted stems, prefixes and suffixes. While the second consists of a lexicon specifying verbal entries in terms of roots and paradigm IDs. By merging these two resources in one model, all possible analyses are provided to cover more than 2600 EMR verbs. Evaluation of CALIMA$_{GLF}$ on 620 verbs gives an accuracy of 81%.

Taji et. al (Taji, et al., 2018) presented CALIMA$_{Star}$ a multi-system that includes an MSA morphological generator. This generator follows concatenative morphology and relies on an extended database of Buckwalter. It contains tables of stems, clitics and compatibility rules that are used in order to avoid generating incorrect words. Taking into consideration only compatible morphemes, the implemented generator expects a lemma and a POS category as input to generate all possible forms.

Torjman and Haddar (Torjmen and Hadder, 2019) automatically built a Tunisian annotated vocabulary containing 150 460 words using finite-state transducers. They started by building a lexicon of 1 452 annotated lemmas and implemented a set of morphological local grammars in NooJ linguistic platform (Silberztein, 2005) following concatenative morphology. Local grammars are then concerted to transducers which govern the concatenation of Tunisian morphemes with these lemmas. To test this vocabulary, they collected 18 134 words from social media and realized that the coverage is 58.5%.

Table 1 presents a summary of various works that target Arabic words generation. By analyzing related information, we notice first that no works targeted the Moroccan Arabic. Additionally, the majority of these works deal with MSA and implement finite-state transducers to generate corresponding vocabularies. In addition, the coverage, by comparing the generated vocabulary with an MSA corpus, in the claimed works reaches an unsatisfying score. In the following sections, we present the Moroccan Arabic morphology and then the followed approach to generate the corresponding vocabulary (MORV).

3. Moroccan Arabic

3.1. General overview

The Moroccan constitution[2] recognizes two official languages: Arabic and Tamazight. Both have their spoken (informal) and written forms and are used in official venues as well as informal situations. The spoken form of Arabic in Morocco is the Moroccan Arabic dialect and it is considered as the mother tongue of Moroccans besides other spoken forms of Tamazight such as Tarifit, Tashelhit and Tamazight (Ennaji, 2005). However, based on the latest available figures[3], most of the Moroccans (91%) can speak Moroccan Arabic, while only 27% of them can use at least one of the spoken forms of Tamazight. Thus, obviously Moroccan Arabic is the primary dialect in Morocco which is mainly used in informal venues such as communication between people and exchanging information.

Recently, with the advent of the Internet and new technologies in Morocco, there has been an outstanding explosion and dispersion of information sources. As Moroccan Arabic is the primary language of communication between Moroccans, this dialect has become dominant in different web sources expressed various forms such as written text, audio and video materials. Consequently, various opportunities are open to better understanding the Moroccan community in different contexts by analyzing and lifting out useful information from the text they produce every day on the web. Within this scope, NLP techniques can be applied to address a wide variety of tasks such as sentiment analysis, topic identification, user's behavior prediction, events detection, to name a few.

According to several linguistic experts such as Ouadghiri (Ouadghiri, 2013), Moroccan Arabic diverges from MSA at the lexical and the phonological levels according to three factors as follow:

- The periodic time: Moroccan Arabic has evolved from its interaction with the Tamazight language in the 7th century to the 20th century with the influence of French and Spanish (during the protectorate period from 1912 to 1956).
- The geographic area: spoken MA in the east of Morocco differs from the MA spoken in Moroccan Sahara[4] and Doukkala[5].
- The speech context: spoken Moroccan Arabic differs according to the context of the speech. For example, in TV programs and education venues, spoken Moroccan Arabic is heavily influenced by MSA where speakers may also alternate between MA and MSA. In other situations, like communication between people and family, spoken Moroccan Arabic can include French words.

This situation poses several problems in MA identification and processing. Hence, since we are not linguistic experts, we are so far from determining and defining standards for Moroccan Arabic. However, as we deal with digital content expressed in MA that presents several business opportunities, we limit the scope of our research on the MA used on the Internet.

3.2. Moroccan Arabic morphology

In this section, we provide an overview of the Moroccan Arabic morphology on which we have relied to generate MORV. In fact, we have based our findings on the works of Moroccan linguistic researchers (Medlaoui Mennabhi, 2019) (Ouadghiri, 2013) (Chafik, 1999). According to these researches, MA morphology is inspired by Arabic morphology with limited exceptions which come from the influence of Tamazight language (Chtatou, 1997). Thus, in the light of these works and our understandings as Moroccan native speakers, we identify main MA words categories and their corresponding morphological attributes. We present also for each category the various rules that can occur during the concatenation of a lemma with affixes and clitics.

As with the Arabic language, Moroccan Arabic (MA) has a rich morphology. In general, MA vocabulary is composed of words that can be classified into three categories: Noun, Verb and Particle. A word can be decomposed to morphemes as described in Fig. 1 where affixes and clitics are used in order to make new words starting from a lemma/stem without changing the POS.

In this paper we define lemma, stem, word, and other morphemes as follows:

- Prefixes: attach before the lemma/stem and states the inflection;
- Suffixes: attach after the lemma/stem stating the inflection;
- Affixes: the set of prefixes and suffixes;
- Proclitics: attach before the lemma/stem and states a syntactic role;
- Enclitics: attach after the lemma/stem and states a syntactic role;

[2] https://www.maroc.ma/en/content/constitution

[3] General Moroccan census performed in 2014 (http://rgph2014.hcp.ma)

[4] Region of Morocco located in the south

[5] Situated in west-central Morocco

Table 1
Arabic morphological generators.

Work	Generator	language	morphology	implementation	size	accuracy
(Beesley, 1996) & (Beesley, 2001)	Xerox	MSA	T + C	FST	72 M	–
(Cavalli-Sforza, et al., 2000)	MORPHE	MSA	T + C	–	–	–
(Habash, 2004)	Aragen	MSA	C	–	–	76%
(Habash et al., 2005), (Habash & Rambow, 2006) and (Habash & Rambow, 2007)	MAGEAD	MSA & LEV	T + C	FST	–	–
(Shaalan et al., 2007)	–	MSA	C	Prolog	–	–
(Attia, et al., 2011) & (Shaalan et al., 2012) & (Attia, et al., 2014)	AraComLex	MSA	C	FST	9 M	–
(Amid Neme, 2013)	–	MSA	C	FST	2.5 M	–
(Doumi, et al., 2016)	–	MSA	T + C	FST	11 M	70%
(Khalifa, et al., 2017)	CALIMA$_{GLF}$	Gulf dialects	T + C	–	–	81%
(Taji, et al., 2018)	CALIMA$_{Star}$	MSA	C	–	–	–
(Torjmen, et al., 2019)	NooJ	Tunisian	C	FST	150 K	59%

C: concatenative morphology
T: Templatic morphology

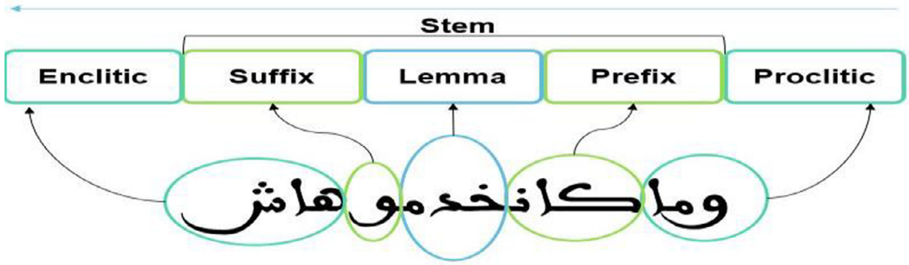

Fig. 1. MA word decomposition.

- Lemma: it is the uninflected base form of a word without affixes and clitics. For verbs, it is conjugated in the perfective, 3rd person and singular form. In the case of nouns and adjectives, the lemma is the singular indefinite form.
- Stem: it is the combination of lemma with affixes.
- Word: can be either a lemma, a stem or the combination of the stem/lemma with clitics (fully inflected form).

Following this definition, we can decompose for example the Moroccan Arabic (MA) word وماكانخدموهاش /And we don’t process it/ (wmakankhdmohach) to several morphemes as illustrated in Fig. 1.

3.3. Verbs

Given that MA is a variant of the Arabic language, not only Arabic lexicon is borrowed but also Arabic grammar rules. Thus, most of them are the same for Moroccan Arabic and in some cases, they are altered in order to meet MA phonology. Accordingly, besides applying Arabic conjugation rules to MA verbs, there are some MA verb conjugations that are slightly different from Arabic. For instance, given the Arabic conjugated verb in the first person at the present tense أكتُبُ /I write/, the Arabic prefix أ is replaced by the MA prefix كان and the last diacritic Damma () is transformed to Soukoun diacritic () to obtain the MA verbكانكتبْ (kanktb). For making negative statements, MA follows a similar pattern to French language by placing the lemma verb between the proclitic ما and the enclitic ش. Passive verbs are obtained by adding the prefix ت to a given verb. Table 2 illustrates some verb conjugation cases in the present tense, the negative state and the passive voice.

Table 2
MA verb conjugation cases.

Conjugation case	MA	Arabic	Meaning
Present tense in the 1st person	كانكتب (kanktb)	أَكْتُبُ	I write
Negation form	مانكتبش (manktbch)	لَنْ أَكْتُبَ	I will not write
Passive voice	تكتبات (tktbat)	كُتِبَتْ	It has been written

In order to facilitate generating MA verbs, we seek to categorize MA verbs according to their common conjugation rules. Thus, one key is considering weak letters in order to categorize MA verbs. In fact, weak verbs (as in Arabic) can also be present in the MA lexicon given that 81% of the MA lexicon is borrowed from Arabic according to a previous study (Tachicart, et al., 2016). A weak verb has one or two weak letters in its root. The letters that make an MSA verb weak are Waw (و), Alif (ا) and Yae (ى). Particularly, given the MA standards spelling adopted in this work, only Waw (و), Alif (ا) are considered weak. In this context, if we consider the orthographic transformations that occur during the concatenation process between morphemes, MA verbs can be categorized into five sets according to the number of letters and the presence of weak letters as illustrated in table 3.

The first set does not undergo any changes in the lemma during the concatenation process. It includes verbs that have no weak letters such as كتب /to write/ (ktb) and زرب /to hurry up/ (zrb) in addition to weak verbs with more than three letters such as تخاصم /to argue/ (tkhasm) and هاجر /to emigrate/ (hajr). The second set is composed of verbs having three letters with the presence of Alif in the middle such as شاف /to see/ (chaf) and قال /to say/ (qal). In the concatenation process, for example in the present tense, the Alif is transformed either to و Waw or to ي Yae such as illustrated in table 3. Additionally, the Alif is deleted in the past tense such as شفت /I saw/ (chft). The third set is composed of verbs that have Alif as the last letter. This letter is transformed in some cases to ي Yae such as the concatenation with present and imperative affixes. In the fourth set, we can find verbs that are composed of two letters with the presence of Chedda () in the last such as شمّ /to sniff/ (chmm) and سدّ /to close/ (sdd). In some cases such as the concatenation with past affixes, Yae (ي) is added to the lemma.

Table 3
MA verbs categories according to concatenation variations.

set	Verb features	Example	Transformation	Meaning
1	- Strong verbs[a] composed of 3 letters - Composed of more than 3 letters given that Alif is not the end position	كتب (ktb) تخاصم (tkhasm)	ماكاي+كتب+وش ماكاي+تخاصم+وش	They don't write They don't argue
2	Composed of three letters where Alif is the second position	شاف (chaf) مال (mal)	كاي+شوف+ونا كاي+ميل+ها	He sees us he leans in
3	The Alif letter is the end position	سطاسيونا (stasiona) مشا (mcha)	كاي+سطاسوني+ها كاي+مشي	He parks it He goes
4	Composed of two letters with chedda () at the end position	سدّ (sdd) شمّ (chmm)	سدّي+ناها شمّي+ت	We closed it I sniffed
5	Irregular verbs[b]	خدا (khda) كلا (kla)	كي+اخد ماي+اكل+وش	He takes They don't eat

[a]Strong verbs do not have weak letters.
[b]Regular verbs set (from set 1 to set 4) are conjugated according to rules that the large majority of verbs in the language use. While irregular verbs (set 5) are conjugated according to different rules.

The fifth set gathers a few irregular verbs such as خدا /to take/ (khda) and كلا /to eat/ (kla) where the Alif may change its position to the first letter. For example, in the case of conjugating كلا to the present tense with the third person ياكل /he eats/ (yakl). It should be noted that contrary to MSA, Moroccan Arabic lexicon lemmas does not include verbs where Yae or Waw is the end position and consequently, only Alif can be that position.

3.4. Nouns

The majority of the MA lexicon (67%) is composed of nouns according to previous work (Tachicart, et al., 2014). In this work, we fit Arabic standards to MA nouns categorization used in (Jaafar and Bouzoubaa, 2015) and thus we decompose MA nouns to several types in order to prepare the necessary rules for the generation process. In this context, we consider noun types that are mentioned in table 4 where each type is compatible with a specific morpheme set. We illustrate each noun category with an example in order to understand our classification. For example, pronouns are compatible only with negation clitics and the conjunction و. Additionally, adverbs are not compatible with definite clitics such as ال' whereas common nouns and adjectives are compatible with almost all nominal morphemes.

3.5. Particles

Particles are words to which noun and verb symptoms cannot apply (Namly, et al., 2016). Contrary to nouns and verbs, particles cannot be inflected. However, they can be concatenated with some morphemes. We consider five types of Moroccan particles as mentioned in table 5: interjections, prepositions, interjections, conjunctions, exceptions and interrogations. Each category can be concatenated with some morphemes as exhibited in table 6.

Table 4
MA nouns categories.

Category	Example	English equivalent
Common	سكات	Silence
Adverb	تقريبا	Approximately
Adjective	فقير	Poor
Pronoun	نتوما	You
Proper	المغرب	Morocco
Number	جوج	two
Broken plural	بيبان	doors

Table 5
MA particles categories.

Category	Example	English equivalent
Interjection	اوّاه	Oh!
Preposition	تحت	under
Conjunction	و	and
Exception	غير	except
Interrogation	شنو	what

3.6. Morphological attributes definition

At the morphological level, most of the Moroccan rules are extended from Arabic since Moroccan Arabic is a variety of Arabic language. In this context, we standardize MORV morphological information according to the ALESCO[6] standards for Arabic morphological analyzers.

In tables 7 and 8 below, we detail MORV morphological information. Indeed, MORV considers the same Arabic morphological categories (noun, verb and particle) where each category can be assigned the following attributes:

- Gender: Verbs can be separated into three classes: feminine and masculine and common. Gender does not apply to particles.
- Number: refers to the quantity of countable nouns or to the number of verb-subject.
- Tense: is the time described by a verb which can be in the past, the present, the future or the imperative.
- Person: refers to someone taking part in the event which is expressed by a verb. It can be with assigned three values: the first, the second or the third.

[6] http://www.alecso.org/

Table 6
MA particles compatibility.

Category	M1	M2	Example	English equivalent
Interjection	Not compatible	Not compatible	وا	Oh !
Preposition	Partially compatible	Partially compatible	وتحتها	and under it
Conjunction	Not compatible	Not compatible	و	and
Exception	Partially compatible	Partially compatible	وغيرها	and others
Interrogation	Partially compatible	Not compatible	واشنو	and what

M1: Morphemes taking place before lemma
M2: Morphemes taking place after lemma

Table 7
MORV morphological information - Common features.

Morphological category	Associated attributes	Possible values
Verbs And Nouns	root	Indefinite but limited
	lemma	Indefinite but limited
	gender	masculine; feminine; common
	number	singular; plural
	form	affirmative; negative

Table 8
MORV morphological information - Specific features.

Morphological category	Associated attributes	Possible values
Verbs	tense	present; future; past; imperative
	person	1; 2 ; 3
	voice	active; passive
	transitivity	yes; no
Nouns	state	definite; indefinite; not applicable
Particles	negation	1;0

- Voice: In a given sentence, it describes the relationship between the subject and the verb. There two verb voices: the active and the passive.
- Transitivity: a verb that accepts one or more objects is transitive.
- State: a noun is indefinite when it is unspecific. By adding the prefix ال, the word state is then transformed to definite.
- Form: negation can be applied to words in affirmative form by using the affixes ما and ش and thus the word form is transformed to negative.

4. Methodology

Following a rule-based approach, generating MA words involves the application of a set of linguistic and orthographic rules that define how morphemes can be concatenated to each other. These rules vary from one-word category to another (verbs, nouns, particles). As stated in section 1, concatenative morphology and templatic morphology are two directions to follow in order to generate a MA word. In this section, we describe the design and the implementation of our generator MORG.

4.1. Design

As illustrated in Fig. 2, four key components enable MORG:

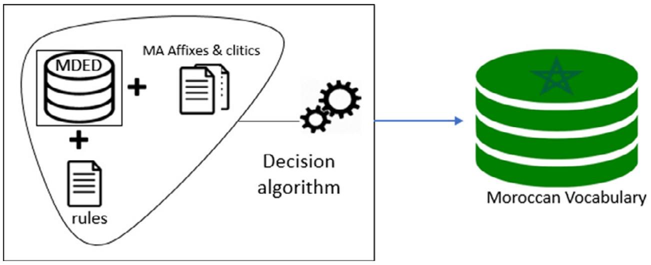

Fig. 2. MA word decomposition.

- Lexicon with labelled lemma named MDED;
- Labelled morphemes table containing both affixes and clitics;
- Rules and constraints governing the spelling and concatenation of morphemes and lemmas;
- Decision algorithm that combines the previous items.

Indeed, the design of the MORG system (that enables the generation of the Moroccan vocabulary 'MORV') is flexible enough to be extended easily. In fact, the used resources and the generation rules are designed in a generic format and are stored in separate tables that allow not only efficient management of both resources and rules but also better maintainability and scalability. Consequently, languages supporting the concatenative morphology can be accommodated seamlessly especially Arabic dialects that share a lot in common with the Moroccan Arabic dialect.

4.1.1. Lexicon

First, as an MA lexicon, we used the Moroccan Dialect Electronic Dictionary (MDED) built in previous work (Tachicart, et al., 2014). To the best of our knowledge, it is the most comprehensive Electronic lexicon for MA that is periodically updated. It contains almost 12,000 MA entries written in Arabic letters and translated to MSA. In addition, one major MDED feature is the annotation of its entries with useful metadata such as POS, origin and root as illustrated in Table 9. For instance, the MA noun ماكلة /food/ is originated from MSA with the root كلا.

4.1.2. MA morphemes table

Besides the lexicon, the morphemes table is a central resource for our morphological generator. As far as we know, there is no work that gathers MA morphemes and exhibits their features. Indeed, 402 MA affixes and clitics were manually created and linguistically checked. Morphemes table is composed of 24 atomic affixes, 43 atomic clitics and 335 compound morphemes. The main advantage of this table is its rich morphological information such as POS, negation, and personas as illustrated in Table 10. For example, the morpheme وكان is composed of the prefix كان and the clitic و. It is compatible with verbs in the present tense, the first person, plural form and all genders. The negation does not apply to this morpheme. Table 11 presents also some insights about the Moroccan morphemes table.

Table 9
Sample of MDED lexicon.

MA	MSA	POS	root	origin	English translation
ماكلة	طعام	Noun	كلا	MSA	food
شحال	كم	particle	شحال	MSA	How much
سطاسيونا	ركن	Verb	سطاسيونا	French	To park

Table 10
Sample of the MA affixes and clitics.

morpheme	value	composition	pos	tense	pers	neg	num	gen
clitic	و	atomic	verb	all	all	all	all	all
prefix	وكان	و+كان	verb	present	1	0	p	all
prefix	بال	ب+ال	noun	–	–	0	all	all
suffix	ين	atomic	noun	–	–	0	p	all
proclitic	وماب	و+ما+ب	noun	–	–	1	all	all
suffix	ات	atomic	verb	all	3	0	s	f
enclitic	كش	ك+ش	verb	all	all	1	s	all
clitic	و	atomic	verb	all	all	all	all	all

Table 11
Distribution of MA affixes and clitics According to POS.

Type	M1	M2	Total	Percentage
nominal	23	67	90	20,81%
verbal	167	145	312	79,19%
Total	200	202	402	100%
nominal	23	67	90	20,81%

4.1.3. Concatenation rules

After preparing the lexical resources that describe different MA morphemes, it is necessary to define rules and constraints governing the concatenation of these morphemes in order to form new MA words, then build and implement the decision algorithm. Therefore, morphological and orthographic rules are stored in three separate tables. Adding new rules or updating them can be performed easily and does not affect MORG overall performance. The first table gathers the morphological attributes of morphemes concatenation. The second indicates which morpheme can be concatenated with which other and in which order. The third specifies orthographic adjustments required in order to convert the generated word into a correct spelling.

4.1.3.1. Morphological attributes table. Regarding the generated word, defining the value of each morphological attribute such as the person, gender, number, etc. relies on the morphological information of each morpheme composing this word. Table 12 shows the effect of combining morphemes on the value of the word morphological attribute. For example, the third line indicates that combining (inside a verb word) a prefix in the present tense with a suffix that accepts all tenses should produce a verb in the present tense.

4.1.3.2. Compatibility table. In order to avoid obtaining impossible words such as والشربنا /and the we drink/ we build the compatibility table. This table is hand-written and determines for each lemma category which morpheme preceding the lemma (proclitic or prefix) can be concatenated with another morpheme that is placed after the lemma (enclitic or suffix) inside a word. To build this table, we start by assuming all morphemes are compatible with each other and thus we generate the corresponding list. Then, we manually checked and excluded morphemes combinations that can produce an incorrect word. For example, even if the prefix ال and the suffix كم are compatible with nouns, they cannot be concatenated together in the same word. As an illustration, concatenating the previous morphemes (ال and كم) with the lemma طيارة /plane/ produces the incorrect word الطيارتكم /the your plane/. As a result, this morphemes combination is excluded from the compatibility table. Table 13 presents a sample of the compatibility table.

4.1.3.3. Orthographic adjustments table. Related information of the constraints governing lemmas concatenation with other morphemes is held in a separate table. Given that some morphemes boundaries are affected during the concatenation process; some orthographic adjustments should be performed in order to correct the generated word. As illustrated in table 14, the newly generated word (which is an intermediate representation) تايتتمشّا /He walks/ arises from the concatenation of the prefix تايت with the lemma تمشّا. However, it presents an orthographic imperfection consisting in a double ت letter. Thus, one ت letter should be deleted in order to produce the correct word form تايتمشّا. The same table illustrates other orthographic adjustments that may occur after combining morphemes.

4.2. Implementation

Rules to generate MORV are implemented using Finite State Transducers (FSTs). These machines have been used in various NLP applications and show their capacity to model different NLP fields such as generation, analysis and speech recognition, etc. as cited in (Karttunen, 2000) and (Mohri, 1996). In fact, an FST is an enhanced finite-state automaton (FSA). While FSA can only accept or reject a string, FST is more general given that it produces output string as well as reading input by defining relations between them; Our FSTs consist of a finite number of states (listed in table 15 and illustrated in Fig. 3) which are linked by transitions and labelled with an input/output pair. In the following figure, we define morpheme boundaries with the (^) mark and word boundaries with the (#) mark. As an example, we consider words generation that takes as input verbal lemmas and all types of morphemes (taking place before and after the lemma). Hence, reading a morpheme taking place before the lemma (M1) leads to the q1 state and so on, until the final state q4 that produces the generated word with corresponding morphological attributes.

In order to build MORV, a finite-state network in cascade is created by defining two levels of morphology. The highest level hosts the lexical string corresponding to the combination of different morphemes and lemma with their corresponding tags. As illustrated in Fig. 4 that exhibits the creation of the word مامشاتش /

Table 12
Morphological attributes table.

Morphological attribute	M1	Lemma	M2	Resulted attribute
gender	feminine	verb	–	feminine
person	all	verb	all	all
tense	present	verb	all	all
number	all	verb	–	singular
person	all	noun	all	all

Table 13
Compatibility table.

Lemma	M1	M2	Example in a word	English equivalent
verb	ماكان	ش	ماكانخدموش	We don't work
verb	غي	وهم	غيعرضوهم	They will invite them
noun	ال	ات	البيكالات	bicycles
noun	ب	تنا	بطوموبيلتنا	with our car
particle	وما	كش	ماعندكش	You don't have

Table 14
Orthographic adjustments examples.

Concatenation	Intermediate representation	Corrected form	English
تايت+تمشّا	تايتتمشّا	تايتمشّا	He walks
مشا+ات	مشاات	مشات	She leaves

Table 15
FST states and transitions.

states	M1	M1 + lemma	M1 + lemma + M2	Intermediate Representation + tags
q0	q1	–	–	–
q1	–	q2	–	–
q2	–	–	q3	–
q3	–	–	–	q4

she didn't leave/, these elements are chained together with boundary markers (+) and present input for the first set of transducers (FST1). The latter maps each lexical string to an intermediate representation which may require some orthographic adjustments in order to meet the Moroccan Arabic spelling constraints. Thus, the intermediate form presents, in turn, an input for the second set of transducers (FST2) that maps this combination form to the orthographically correct surface form.

Finally, the last task consists of compiling MORV and building a vocabulary of Moroccan words annotated with different morphological attributes. To this end, we use the previous finite-state network to create separately three lexical databases: the first gathers nouns (including irregular forms such as broken plural), the second includes verbs while the third is composed of particles. The reason why MORV compilation is performed separately according to the lemmas category is that the concatenation rules and also the orthographic adjustments differ accordingly. Table 16 presents a global insight about MORV content while table 17 illustrates a sample of generated forms regarding the verb سطاسيونا /to park/. Additionally, an extended sample of MORV containing further entries can be found at the *SAFAR website*[7].

5. Evaluation and discussion

5.1. Quantitative evaluation (coverage)

In order to evaluate the MA generated vocabulary regarding the coverage of Moroccan Arabic, we check whether the used Moroccan Arabic words orthographically exist in MORV. Thereby, this evaluation does not concern the associated annotations with MORV generated words. To this purpose, we use a test corpus extracted from the MA User-generated Text (UGT) which is introduced in previous work (Tachicart and Bouzoubaa, 2019). Our test corpus[8] is manually normalized following the same MORV orthographic rules and includes 1000 MA sentences containing 10,564 words. In fact, the manual task that should be performed in order to prepare such a corpus impedes us to increase its size. The goal of the evaluation is to ensure that each UGT word in the test corpus is orthographically recognized using MORV. Obtained results show that the percentage of the recognized words is 84% which means that only 16% of the test corpus words are missed. This score is encouraging if we compare MORV to MSA vocabularies which do not exceed 87% in the language coverage such as BAMA and AraComLex. In table 18 below, we detail the quantitative evaluation by providing the coverage in addition to the out of vocabulary (OOV) rates according to the POS.

By examining the OOV words list, we notice that (without considering named entities) most of the words belong to the MSA vocabulary or completely have an MSA morphology due to the use of code-switching in MA texts. For example, the wordsتنسجم /fit into/, الأقمصة /shirts/, and الاستهجان /boos/ are MSA words that are included in the MA evaluated sentences. Besides, there are some examined cases in the OOV list where the items can be considered as MA words such as شيباهم (he shipped the items). This can be explained by the fact that MDED lexicon does not include the corresponding lemma which could be considered as a new word for Moroccan Arabic.

5.2. Qualitative evaluation (performance rates)

Building new NLP resources such as MORV is very important to various NLP tasks. The importance of such resources depends not only on their size and coverage but also on the credibility of the provided information. In this perspective, we assess MORV performance using regular evaluation metrics namely: precision, recall, accuracy and F-measure. These metrics are calculated using the regular parameters True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as described in table 19.

Based on these parameters, for each input word, we define the precision as the number of correct annotations found in MORV compared to the total number of the annotations that correspond

[7] http://arabic.emi.ac.ma/morv.

[8] Available at http://arabic.emi.ac.ma/morv.

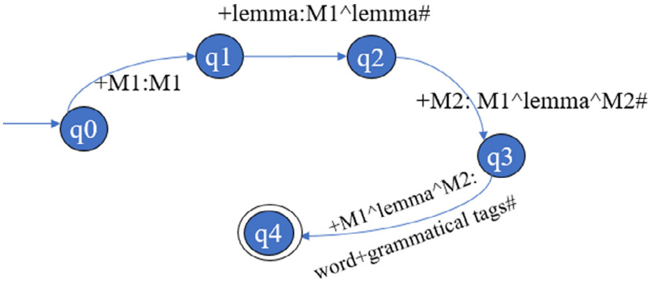

Fig. 3. FST transitions (handling verbs).

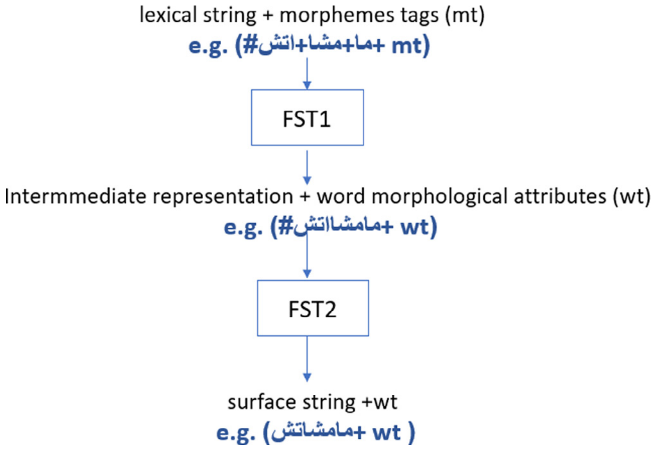
lexical string + morphemes tags (mt)
e.g. (#ما+مشا+اتش+ mt)
FST1
Intermmediate representation + word morphological attributes (wt)
e.g. (#مامشااتش+ wt)
FST2
surface string +wt
e.g. (مامشاتش+ wt)

Fig. 4. FST1 handling a set of Moroccan verbs.

to the input word in MORV. Alternatively, it can be calculated as in the following formula:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Moreover, the recall is defined as the number of correct annotations found in MORV compared to the expected annotations that are correct and should be found. The recall can be calculated by the following formula:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Also, accuracy expresses the proportion of false annotations. It is calculated as in the following formula:

$$Accuracy = \frac{TP}{TP + FP + FN} \quad (3)$$

Finally, the F-measure is defined as the harmonic mean of precision and recall as follow:

$$Fmeasure = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Yet, before engaging in the qualitative evaluation, it is necessary to prepare a manually annotated test corpus. To this end, we selected carefully 25 sentences containing 304 MA words from the above test corpus (used in the quantitative evaluation) and asked a linguist expert to provide for each word all possible annotations without taking into consideration the context and also considering the same morphological attributes provided by MORV. Then, we extracted MORV associated annotations for each word (in the new test corpus) and perform a comparison against that in the test corpus regarding all morphological attributes. Table 20 below provides a summary of the evaluation process.

Given the obtained results, we observe first that the precision rate is 94,88% which expresses the correctness of the annotations that are associated with MORV words. It relates in fact to the low false-positive rate including correct annotations at the morphological level but incorrect at the semantic level such as the word كانسكنوكم /we live in you/. We are not impressed by the high precision rate since the generation process follows a rule-based approach where morphological rules are checked by linguist experts. Besides, we obtained a recall of 81,42% which means that not all relevant results are returned but it can be relatively considered a good score. The reason behind missing some relevant results is that contrary to the precision calculation, we considered the OOV words in the recall calculation. This increased the number of associated annotations that should be found in the test corpus. In light of the evaluation rates, it is clear that including new lemmas in MDED lexicon and running the corresponding generation process should decrease MORV OOV rate and thus it is a key factor towards maximizing the recall. Regarding accuracy and F-measure rates, we obtained respectively 77.99% and 87,64%. In fact, the cost of false positives and false negatives are very different in the case of evaluating MORV associated annotations. Moreover, these data are not symmetric where values of false positive and false negatives are almost the same. For this reason, it is useful to look both at precision and recall as metrics of MORV evaluation.

To compare MORV to the other existing works, we consider only Arabic dialects. Thus, we believe that MORV has shown the largest size and the best precision as reviewed in Table 1 and reported in the quantitative and the qualitative evaluation. As an example, while the Tunisian vocabulary contains 150 k forms generated from 1452 lemmas, MORV exploit 12.000 lemmas to generate 4.68 M forms. Additionally, the accuracy of MORV outperforms the claimed accuracy of the other Arabic dialects vocabularies.

It should also be noted that all experiments that enabled the building and evaluating MORV were performed on a workstation having the following characteristics: CPU = i7 @ 2.7 GHz 2.7 GHz, RAM = 32GO, Operating System = Win10, 64 bits.

Typically, a strategic requirement for research and development in the NLP field is the creation of high-quality language resources given that the performance of NLP tools usually relies on the quality of these resources. Therefore, we believe that extending MORV or creating new resources will pave the way towards addressing Moroccan NLP tasks. For this reason, it would be useful to follow the templatic morphology that involves the creation of MA roots,

Table 16
MORV general insights.

Morphological category	Lexicon entries	Lexicon percentage	MORV generated forms	MORV percentage
Verbs	3130	26%	2.021.152	43,15%
Nouns	8598	73%	2.655.460	56,68%
Particles	118	1%	8.154	0,17%
Total	12.000	100%	4.684.766	100%
Verbs	3130	26%	2.021.152	43,15%

Table 17
Sample of the MA affixes and clitics.

word	lemma	transitivity	root	suffix	prefix	form	tense	number	gender	person
ماكاتسطاسيونيش(You don't park)	سطاسيونا	yes	سطسين	ش	ماكات	1	2	s	m	2
ماكاتسطاسيونيش(You don't park)	سطاسيونا	yes	سطسين	ش	ماكات	1	2	s	f	3
وماغتّسطاسيوناش(And it will be not parked)	سطاسيونا	yes	سطسين	ش	وماغتّ	1	3	s	m	2

Table 18
MORV quantitative evaluation according to POS.

Morphological category	Verbs	Nouns	Particles	Total	Average
Number of words	3861	5977	2518	12,356	–
Coverage	93%	86%	97%	–	84%
OOV	9%	16%	3%	–	16%

Table 19
MORV qualitative evaluation rates.

	Positive	Negative
True	Correct annotation (corresponding to the input word) found in MORV	Incorrect annotation (corresponding to the input word) not found in MORV
False	Incorrect annotation (corresponding to the input word) found in MORV	Correct annotation (corresponding to the input word) not found in MORV

Table 20
MORV qualitative evaluation rates.

Morphological category	Verbs
Associated annotations	255
Precision	94,88%
Recall	81,42%
Accuracy	77,99%
F-measure	87,64%

patterns and the rules governing the use of roots and patterns to create new MA words. This may help in comparing the results of the concatenative morphology that has already been covered in this work and the templatic morphology.

6. Conclusion

In this paper, we have presented MORV the first NLP resource that targets Moroccan morphology. MORV is a flexible, FST-based tool for automatically generating Moroccan Arabic dialectal words. The FST topology is inspired by earlier approaches used to generate Arabic words. Additionally, we have exploited Moroccan Arabic resources for this purpose and generated 4.68 M Moroccan words with full morphological attributes. We have also shown how important keeping the linguistic knowledge separated from the processing algorithm in order to ensure MORV extensibility. MORV evaluation shows that it reaches a precision of 94% and covers 84% of the Moroccan Arabic used in social media. These rates can be improved by incorporating new lexicon entries and run the corresponding generation process which is the subject of our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al-Shargi, F., Kaplan, A., Esk, er, R., Habash, N., Rambow, O., 2016. Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic. European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1300–1306.

Altantawy, M., Habash, N., Rambow, O., Saleh, I., 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. European Language Resources Association (ELRA), Valletta, Malta.

Amid Neme, A., 2013. A fully inflected Arabic verb resource constructed constructed from a lexicon of lemmas by using finite-state transducers. Revue de l'Information Scientifique et Technique, 1–13.

Attia, M., Choukri, K., Yaseen, M. (2005). Specifications of the Arabic Written Corpus produced within the NEMLAR project.

Attia, M., Pecina, P., Toral, A., van Genabith, J., 2014. A corpus-based finite-state morphological toolkit for contemporary arabic. Journal of Logic and Computation 24 (2), 455–472.

Attia, M., Pecina, P., Toral, A., Tounsi, L., van Genabith, J. (2011). An Open-Source Finite State Morphological Transducer for Modern Standard Arabic. 9th International Workshop on Finite State Methods and Natural Language Processing. Blois, France

Bauer, A., Hoedoro, N., Schneider, A. (2015). Rule-based approach to text generation in natural language (ATML3). the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data, the Special Industry Track and the RuleML 2015 Doctoral Consortium hosted by the 9th International Web Rule Symposium (RuleML 2015). Berlin, Germany: CEUR-WS.org.

Beesley, K.R., 1996. Arabic Finite-State Morphological Analysis and Generation. ACL, Copenhagen, Denmark, pp. 89–94.

Beesley, K. R. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. the ACL 2001 Workshop on Arabic Language Processing: Status and Prospects. 1, pp. 1-8. WS, USA: ACL.

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

Cavalli-Sforza, V., Soudi, A., Mitamura, T. (2000). Arabic Morphology Generation Using a Concatenative Strategy. 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, Washington, USA: Association for Computational Linguistics.

Chafik, M., 1999. The Moroccan dialect: a field of confluence between Amazigh and Arabic Languages. Academy of the Kingdom of Morocco, Rabat.

Chtatou, M., 1997. The influence of the Berber language on Moroccan Arabic. International Journal of the Sociology of Language 123 (1), 101–118.

Diab, M., Habash, N., Rambow, O., Altantawy, M., Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. In Lrec workshop on semitic language processing. LREC Workshop on Semitic Language Processing, (pp. 66-74). Valletta, Malta

Doumi, N., Lehireche, A., Maurel, D., Abdelali, A., 2016. A Semi-automatic and Low Cost Approach to Build Scalable Lemma-based Lexical Resources for Arabic Verbs. International Journal of Information Technology and Computer Science 8 (2), 1–13.

Dusek, O., Jurcícek, F., 2013. Robust multilingual statistical morphological generation models. Association for Computational Linguistics, Sofia, Bulgaria, pp. 158–164.

Ennaji, M., 2005. Chapter Four: Berber Multilingualism, Cultural Identity, and Education in Morocco. Springer, p. 79.

Faruqui, M., Tsvetkov, Y., Neubig, G., Dyer, C., 2016. Morphological Inflection Generation Using Character Sequence to Sequence Learning. the, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp. 634–643.

Habash, N., 2004. Large Scale Lexeme Based Arabic Morphological Generation. Traitement Automatique du Langage Naturel, Fes, Morocco.

Habash, N., 2010. Introduction to Arabic Natural Language Processing. Morgan Claypool Publishers.

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., Rambow, O. (2007). Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. International Symposium on Computer and Arabic Language (ISCAL). Riyadh, Saudi Arabia

Habash, N., Rambow, O., Kiraz, G. (2005). Morphological Analysis and Generation for Arabic Dialects. the ACL Workshop on Computational Approaches to Semitic Languages (pp. 17-24). Ann Arbor, Michigan: Association for Computational Linguistics.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Jaafar, Y., Bouzoubaa, K. (2015). Arabic Natural Language Processing from Software Engineering to Complex Pipeline. First International Conference on Arabic Computational Linguistics (ACLING'15), 19. Cairo, Egypt.

Jinxi, X., 2002. UN Parallel Text (Arabic-English). Linguistic Data Consortium, University of Pennsylvania.

Jisha, P., R, R.R., Rajendran, S., 2011. Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation. International Journal of Computer Applications 13 (8), 15–18.

Karttunen, L. (2000). Applications of Finite-State Transducers in Natural Language Processing. 5th International Conference on Implementation and Application of Automata. London, Ontario, Canada

Khalifa, S., Hassan, S., Habash, N., 2017. A Morphological Analyzer for Gulf Arabic Verbs. the Third Arabic Natural Language Processing Workshop. ACL, Valencia, Spain, pp. 35–45.

Leavitt, John R.R., 1992. MORPHE: A Practical Compiler for Reversible Morphology Rules. Third Conference on Applied Natural Language Processing. Trento, Italy, 233–234. https://doi.org/10.3115/974499.974543.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., Tabessi, D., 2006. Developing and Using a Pilot Dialectal Arabic Treebank. European Language Resources Association (ELRA), Genoa, Italy.

Medlaoui Mennabhi, M. (2019). العربية الدارجة. املائية ونحو: الأصوات، الصرف، التركيب، المعجم. Darija Developpement Centre - Zagora.

Mohri, M. (1996). On some applications of finite-state automata theory to natural language processing. Natural Language Engineering, 2(1), 61 - 80.

Namly, D., Bouzoubaa, K., Youssef, T., Khamar, H. (2016). Development of Arabic particles lexicon. Colloque pour les Etudiants Chercheurs en Traitement Automatique du Langage Naturel. Sousse, Tunisia.

Ouadghiri, A. (2013). لغة الأمة ولغة الأم عن واقع اللغة العربية في بيئتها الاجتماعية والثقافية. Beirut: Dar al kotob al ilmiyah - Beirut - Lebanon.

Shaalan, K., Abdel Monem, A., Rafea, A. (2007). Arabic Morphological Generation from Interlingua. International Conference on Intelligent Information Processing (pp. 441-451). Boston, MA, USA: Springer US.

Shaalan, K., Samih, Y., Attia, M. (2012). Arabic Word Generation and Modelling for Spell Checking. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). Istanbul, Turkey.

Silberztein, M. (2005). NooJs dictionaries. Proceedings of LTC, 5, pp. 291–295. Poland.

Soudi, Abdelhadi, Van den Bosch, A., Neumann, G. (2007). Arabic computational morphology: knowledge-based and empirical methods. In Soudi, Abdelhadi, G. Neumann, A. Van den Bosch, Arabic computational morphology (pp. 3-14). Dordrecht: Springer.

Tachicart, R., Bouzoubaa, K. (2019). An Empirical Analysis of Moroccan Dialectal User-Generated Text. 11th International Conference Computational Collective Intelligence (ICCCI'19). Hendaye, France.

Tachicart, R., Bouzoubaa, K., Jaafar, H. (2014). Building a Moroccan dialect electronic Dictionnary (MDED). 5th International Conference on Arabic Language ProcSessing 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.

Taji, D., Khalifa, S., Obeid, O., Eryani, F., Habash, N. (2018). An Arabic Morphological Analyzer and Generator with Copious Features. the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, (pp. 140-150). Brussels, Belgium.

Torjmen, R., Haddar, K. (2019). Construction of Morphological Grammars for the Tunisian Dialect. In Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications - 12th International Conference, NooJ 2018, Palermo, Italy, June 20-22, 2018, Revised Selected Papers (pp. 62-74). Cham: Springer International Publishing.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer.

Viks, U. (2000). Tools for the Generation of Morphological Entries in Dictionaries. the Second International Conference on Language Resources and Evaluation (LREC'00). Athens, Greece: European Language Resources Association (ELRA).