Lang Resources & Evaluation (2020) 54:1079–1142
https://doi.org/10.1007/s10579-020-09490-9

SURVEY

Language resources for Maghrebi Arabic dialects' NLP: a survey

Jihene Younes[1] · Emna Souissi[2] · Hadhemi Achour[1] · Ahmed Ferchichi[1]

Published online: 25 April 2020

Abstract Diglossia is one of the main characteristics of Arabic language. In Arab countries, there are three forms of Arabic that co-exist: Classical Arabic (CA) which is mainly used in the Quran and in several classical literary texts, Modern Standard Arabic (MSA) that descends from CA and used as official language, and various regional colloquial varieties of Arabic that are usually referred to as Arabic dialects (AD). Deemed to be amongst low-resource languages, these dialects have aroused increased interest among the NLP community in recent years. Indeed, the various Arabic dialects are increasingly used on the social web and may be transcribed in both the Arabic and the Latin script. The latter is known as Arabizi and seems to be more frequently used for some of them. The AD NLP raises many challenges and requires the availability of large and appropriate language resources. In this study, we focus, in particular, on the Maghrebi Arabic dialects (MADs). We propose a thorough review of the language resources (LRs) that have been generated by the various work carried out on the MAD language processing. A survey of the currently online available MAD NLP dedicated-LRs is also compiled and discussed. LRs investigated in this work are essentially data-resources such as primary and annotated corpora, lexica, dictionaries, ontologies, etc.

✉ Jihene Younes
jihene.younes@gmail.com

Emna Souissi
emna.souissi@ensit.rnu.tn

Hadhemi Achour
hadhemi.achour@isg.u-tunis.tn

Ahmed Ferchichi
Ahmed.Ferchichi@gmail.com

[1] Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000 Le Bardo, Tunisia

[2] Université de Tunis, ENSIT, 1008 Montfleury, Tunisia

Keywords Maghrebi Arabic dialects · Language resources · Corpus · Lexicon · Ontology · Natural language processing

1 Introduction

Constructing and evaluating spoken or written Natural Language Processing (NLP) algorithms and systems require the availability of different kinds of resources related to the treated languages and usually referred to as Language Resources (LRs). According to the European Language Resources Association,[1] the term LR refers to a set of speech or language data and descriptions that are accessible in an electronic form and useful for developing or evaluating natural language or speech algorithms and systems. We should, however, note that the definition assigned to the term LR can vary among scholars using it. It ranges from a broad definition encompassing various kinds of data (such as corpora, lexica, thesauri, ontologies, etc.) and tools generating new data and descriptions (such as morphological analysers, taggers, parsers, etc.) (Witt et al. 2009), to a narrower definition where the term LR designates data-only resources (Cunningham et al. 2009). When it comes to data-only LRs, a common classification is to divide these resources into primary and secondary (or derived) resources (Rosner 2009). Primary resources refer to raw data obtained from different textual sources while secondary or derived data refer to data which have been annotated by additional information such as different levels of linguistic descriptions.

Having such different kinds of LRs is crucial for any work aiming at a language study or analysis (El-Haj et al. 2014). Their construction, representation, maintenance and evaluation are important issues in the NLP field and especially for those working on statistical and machine learning methods requiring large scale resources. While a substantial work has been carried out on this area for languages such as English and some other European languages and has led to achieve significant technological advances in language and speech processing, efforts are still required regarding under-resourced languages in order to build various kinds of LRs allowing their automatic processing.

The concept of under-resourced language may have different designations (such as low-resource language, poorly endowed language, etc.) and various definitions. According to Besacier et al. (2013), an under-resourced language refers to "*a language with some of (if not all) the following aspects: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists,* etc.". Other definitions are more centred on Human Language Technologies (HLT) such as that provided by the LORELEI[2] project, considering a low-resource language as a language "*for which*

[1] http://www.elra.info/en/about/what-language-resource/.

[2] LORELEI (Low Resource Language for Emergent Incidents) is a project funded by US government. Its goal is to develop HLT for low-resource languages, in support of missions related to emergent incidents.

no automated human language technology capability exists". As for Duong (2017), this definition is related to a particular NLP task, by considering a language as low-resource for a given task, if there are no existing solutions using available data and performing this task with an adequate performance.

It can be clearly drawn from the above definitions, that a given language may be considered as under resourced when it lacks LRs required to develop one or more NLP tasks involving that language. In this sense, Arabic dialects in general, are deemed to be amongst low-resource languages, in view of the current availability of NLP data resources and tools associated to these dialects (Hamdi et al. 2015; Novotney et al. 2016; Harrat and Meftouh 2017a). It should however be noted that in recent years, Arabic dialects have aroused increased interest among the NLP community. This interest may be explained by, among other reasons, the growing use of these dialects by Arab Internet users, especially on the social web (Zaidan and Callison-Burch 2014; Younes et al. 2015; Samih and Maier 2016a; Alshutayri and Atwell 2017).

Arabic dialects are spoken by more than 440 million people[3] in a region covering Arabia (Arabian Peninsula), North Africa and the Middle East. The classification to which subscribe several researchers (Embarki 2008) is that of Versteegh (Versteegh 1997), which divides modern Arabic dialects into five major dialectal areas, from East to West: (1) *Arabian Peninsula dialects* include the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, Oman and Yemen; (2) *Mesopotamian dialects* include the dialects of Iraq; (3) *Levantine dialects* include the dialects of Lebanon, Syria, Jordan and Palestine; (4) *Egyptian dialects* cover the dialects of the Nile valley: Egypt and Sudan; (5) *Maghrebi dialects* cover the dialects of Mauritania, Morocco, Algeria, Tunisia and Libya.

This study focuses on the Maghrebi Arabic dialects (MADs) in particular. Indeed, MADs have in common the same sociolinguistic variation and are characterized by many common linguistic specificities: the coexistence of several languages (MSA, dialectal Arabic, Berber and French), the influence of French language in written and oral use, the great increase use on social media, the writing with Latin letters, etc. All these features have generated common difficulties and challenges to be overcome when trying to process these dialects. That's why, many research works bring together these dialects. Although most of the work carried out on Arabic dialects has focused on Mashriqi[4] and mainly on Egyptian Arabic (Shoufan and Alameri 2015; Assiri et al. 2015; Harrat and Meftouh 2017b), MADs have been picking increasing interest from NLP researchers in the past few years. As regards surveys, and to our knowledge, only one study proposed by Harrat and Meftouh (2017c) was specifically dedicated to reviewing work on the automatic processing of Maghrebi dialects. Although the survey of Harrat and Meftouh

[3] According to https://data.worldbank.org, consulted in May 2019.

[4] Given the vowel systems and the influence of the Semitic substratum in the East and Berber in the West, the various sets of dialects are grouped into two main classes also based on a geographical distinction (Embarki 2008): (1) Oriental Arabic dialects named *Mashriqi Arabic*, which can be sub-classified into Arabian Peninsula, Mesopotamian, Egyptian and Levantine dialects; (2) Western Arabic dialects named *Maghrebi Arabic*, which can be sub-classified into Mauritanian, Moroccan, Algerian, Tunisian and Libyan dialects.

(2017c) covers a variety of works, it consists in a short and a non-exhaustive review, dealing with only three dialects: Tunisian, Algerian and Moroccan dialects. The authors presented a linguistic overview of the three dialects, then reviewed 39 works (until 2017) according to 7 categories: corpora and lexicons (12 works), identification (5 works), orthography (4 works), morphological analysis (6 works), sentiment analysis (3 works), machine translation (5 works) and other works including sentence boundary detection and diacritics restauration (4 works). Harrat and Meftouh (2017c) provided some information about the followed NLP approaches and sizes of the used corpora. They, however, did not specify the used scripts nor the annotation level of these corpora and didn't provide any information about the availability of MAD language resources.

In the present survey, we aim for a more comprehensive study and a thorough review of the language resources that have been generated by the various work[5] carried out on the MAD language processing. A survey of the currently online available MAD NLP dedicated-LRs is also compiled and discussed. LRs investigated in this work are essentially data-resources such as primary and annotated corpora, lexica, dictionaries, ontologies, etc. Our main goals are to provide a clear picture of what progress has been made towards constructing LRs for the Maghrebi dialects' NLP and their availability to researchers wishing to work in this field.

The reminder of this paper is organized as follows: Sect. 2 is devoted to a general presentation of the MADs. We review their major linguistic specificities on different levels. We also provide some pertinent indications on their current use in the social web and set forth the main difficulties and challenges related the MADs' NLP. In Sect. 3, we identify the constructed MAD primary LRs, mainly, raw text and speech corpora. Section 4 reviews the construction of various annotated corpora (derived LRs). As for Sect. 5, it is dedicated to identifying different kinds of other MAD data-LRs such as lexica, dictionaries and ontologies. We devote Sect. 6 to works that examined the normalization and the codification of the MADs. In Sect. 7, we make a census of the currently online available LRs dedicated to MAD NLP and in Sect. 8, we propose a discussion on the identified resources, supported by numbers and charts illustrating the evolution of works on Maghrebi dialects, the typology of LRs, their breakdown by MAD, etc. Finally, Sect. 9 is dedicated to the conclusion and future perspectives.

2 Maghrebi Arabic dialects (MADs)

2.1 Linguistic specificities

Maghrebi Arabic (or *Maghrebi Darija*) is one of the mother tongues of people of the Maghreb area in North Africa and covers five countries: Mauritania, Morocco, Algeria, Tunisia and Libya. The Maghreb has more than 100 million inhabitants:[6]

[5] This survey examines works that have been published until May 2019.

[6] According to http://countrymeters.info consulted in May 2019.

4.3 million in Mauritania, 36 million in Morocco, 43 million in Algeria, 11.5 million in Tunisia and 6.2 million in Libya. In these countries, at least two mother tongues coexist according to the origin of the inhabitants (Pereira 2005): *Berber*[7] and *Maghrebi Arabic.* In fact, Berber is the mother tongue of nearly 30% of the Maghreb population. Berber speakers represent about 45 to 50% of the population in Morocco, 25 to 30% in Algeria, 1% in Tunisia and between 5 and 10% in Libya and Mauritania. In the other hand, *Maghrebi Arabic* is used by Arabic speakers and Berber speakers, but also between Berber speakers, who do not understand each other, when the Berber variety of each speaker is different.

Speakers of Maghrebi Arabic call their language *Darija* (in Arabic, "دارجة" which is the feminine form of "دارج" meaning "*familiar, common, popular, used*"[8]). *Darija* alludes to colloquial spoken Arabic rather than Modern Standard Arabic (MSA), but it is also common to refer to the Maghrebi Arabic varieties directly as languages. For instance, Moroccan Arabic as *Maghrebi* (Moroccan), Algerian Arabic would be referred as *Dzayri* (Algerian), Tunisian Arabic as *Tounsi* (Tunisian), Libyan Arabic as *Libi* (Libyan) and Mauritanian Arabic as Hassaniya (derived from the name of the Arab tribes called "Beni Hassan").

Maghrebi Arabic is a spontaneous oral language. It is the language in which speakers communicate with each other, but it is also sometimes a literature language and a writing language (Pereira 2005). Indeed, it is a literature language by means of which proverbs, nursery rhymes, tales, riddles, poems are said. It can be a writing language when song lyrics and theatre plays are written in Maghrebi Arabic. Today, it is more used on radio, television and in publicity. In its written form, Maghrebi Arabic is written for a long time mainly with the Arabic script (AS), but it may also be written using the Latin alphabet (LS) [Written content using LS is known as "Arabizi"]. It should be noted that the introduction of new modes of communication (SMS, e-mails, Facebook, Twitter, etc.), widely used in Arab countries, has strengthened dialectal writing, especially in Latin script.

According to Mohand (1999) and Embarki (2008), the Maghrebi Arabic dialects undoubtedly possess common phonetic, morphological, syntactic and lexical features, giving them a particular linguistic character and thereby, clearly differentiating them from Eastern Arabic. The main common characteristics of the MADs include:

- *Mutual intelligibility* The varieties of Maghrebi Arabic have a significant degree of mutual intelligibility, especially between geographically adjacent ones (such as local dialects spoken in Eastern Morocco and Western Algeria or Eastern Algeria and North Tunisia or South Tunisia and Western Libya), but hardly between Moroccan and Tunisian Darija. On the other hand, it is well-known that Maghrebians understand almost all other Arabic dialects.
- *Mixture of many languages* Maghrebi Arabic cannot be understood by Eastern Arabic speakers (from Egypt, Sudan, Levant, Iraq, and Arabian Peninsula) in general as they derive from different substratums and a mixture of many

[7] Berber represents the languages originally spoken by the native populations of the Maghreb prior to their adoption of Arabic.

[8] In Dictionary of meanings "معجم المعاني عربي عربي": http://www.almaany.com.

languages (Mohand 1999; Elimam 2009, 2012; Sayahi 2014; Mzoughi 2015): Punic, Berber, Arabic, Turkish, French, Spanish, Italian and Niger-Congo languages. Sayahi (2014) considers that Berber in particular, has exercised an important influence in the way the Maghrebi dialects formed, distinguishing them from the eastern Arabic dialects at many levels. For these reasons, some linguists like Elimam (2004, 2012), tend to consider Maghrebi Arabic as an independent language.

- *Continual evolution* Maghrebi Arabic continues to evolve by integrating new words. Speakers frequently borrow words from French (in Morocco, Algeria and Tunisia), Spanish (in Morocco) and Italian (in Libya and Tunisia) and conjugate them according to the rules of Arabic with some exceptions (like passive voice for example) (Sayahi 2014; Elimam 2009). Some examples showing the dynamicity of the Tunisian dialect linguistic system are given in Table 1. These examples are also valid for Algerian and Moroccan dialects.
- *Code-switching* Defined as the "*alternate use of two or more languages in the same utterance of conversation*" (Fishman 1999), Code-switching is a result of multilingualism. In the Maghreb, code-switching is frequent and represents a feature of the local way of speech. In text, code-switching generally occurs between MSA and Maghrebi Arabic when it is written in Arabic script and between French/Spanish/English and Maghrebi Arabic when it is written with Latin alphabet. There may also be, but rarer, written texts (especially in the social web and advertising spots) using both Arabic and Latin scripts. Table 2 below shows some examples of these messages. This kind of code-switching adds another difficulty which imposes a reading from left to right of the part written with the Latin letters and in the opposite direction of the Arabic part.
- *Phonological aspects* In Maghrebi dialects, some Non-Arabic phonemes may be used such as/g/ڤ,/p/پ and/v/ڤ. Long vowels are usually shortened, and the three short vowels are often reduced to two. In many cases, CvCC is changed to CCvC (e.g., سَقْف/saqf (*roof*) in Oriental Arabic, is said سقَف/sqaf in Maghrebi Arabic). Also, preference is for final-syllable stress, especially with the reduction of non-stressed short vowels (e.g., كِتَاب/kitaab (*book*) in Oriental Arabic, is said كتَاب/ktaab in Maghrebi Arabic).

Table 1 Words borrowed from French

Dialect words in AS	Dialect Words in LS	French words	Meaning in English
پارتاجا	partaja, partagea, …	*il a partagé*	*He shared*
پيارتاجي	ypartagi, ypartaji, …	*il partage*	*He shares*
ميارتاجي	mpartagi, mpartaji, …	*en situation de partage*	*In situation of sharing*
ميارتاجية	mpartagia, mpartajia, …	*le fait de partager*	*The fact of sharing*
دوش	douch, doch, …	*une douche*	*A shower*
ادوّش، دوّش	adwach, adwech, dawwich, …	*prendre une douche*	*Take a shower*
مدوّش	mdawwich, mdawich, mdewwich, …	*douché*	*Showered*

Table 2 Examples of written MAD using both AS and LS

MAD message	Meaning in English	Source
على خاطر غرامك بالكورة illimité إتمتّع بأنترنات illimité مع les offres internet ابتداء من 26,900د في الشهر fixe Orange وما تراتي حتى ماتش مال CAN	*Because your love for football is unlimited enjoy unlimited internet with Orange fixed internet deals from 26,900 dinars every month and do not miss any CAN matches*	https://www.facebook.com/orange.tn Acceded the June 1, 2019
مراد حاب يديرالـ Geste في عيد الأم مراد ماركا BUT واحد ولّى هو Champion	*Murad wants to make a gesture to Mother's Day* *Murad scored one goal he became champion*	https://www.facebook.com/djezzy Acceded the June 1, 2019
شارك ف Skhawa Dance Challenge هدايا ديال السخاوة في انتظارك	*Participate in Skhawa Dance Challenge gifts waiting for you*	https://www.facebook.com/maroctelecom/ Acceded the June 1, 2019

- *Morphological aspects.* Maghrebi dialects have a specific conjugation distinguishing them from Mashriqi dialects and Modern Standard Arabic. Table 3 below shows some examples.

Although Maghrebi dialects share many common linguistic characters, each one of them has its own linguistic specificities. Table 4 shows some examples of specific characteristics that differentiate them from each other. In this paper, we use MD for Moroccan Dialect, AD for Algerian Dialect, TD for Tunisian Dialect, LD for Libyan Dialect, HD for Mauritanian Dialect; ArD for Arabic dialects; OAD for other Arabic dialects and OL for other languages.

2.2 MAD presence on social networks

As has been the case with the whole Arab world, the Maghreb countries have experienced in the past few years, a greatly increased use of social media. Taking as example Facebook, which remains the most popular social media platform in the Maghreb region, and according to Mourtada and Salem (2014) and Salem (2017), the total number of its users in the Maghreb has almost doubled over the past 3 years, growing from 20.5 in 2014 to 38.3 million users in 2017. On average, Facebook is used in the Maghreb region, by about 38% of its total population.

As for the used languages on the social web, it is important to note that Maghrebi user-generated content is highly multilingual, mainly including the three languages Arabic, French and English. Table 5 gives for each country, a breakdown of users according to the languages they use on Facebook. We can clearly notice the important growth of the Arabic language use in all the Maghreb countries. On average, the percentage of Facebook users that use Arabic, has grown from 43.2% in 2014 to 73.7% in 2016 in the Maghreb region, even though the use of French remains also among the preferences of Maghrebi Facebook users, especially in Tunisia, Algeria and Morocco.

As for the Arabic language used on the social web, it is not limited to MSA, but textual contents expressed in the various Arabic dialects spoken by Maghrebi populations are also proliferating on social platforms (Younes et al. 2015; Samih

et al. 2016; Abidi and Smaili 2017). Moreover, the dialectal productions may be written using both the Arabic and the Latin alphabet. In this regard, Table 6 gives some key figures on the use of dialectal Arabic and its transcriptions in the Algerian and Tunisian social web according to, respectively Abidi and Smaili (2017) and Younes et al. (2015). This table shows the significant use of colloquial Arabic on the web, for both Tunisia and Algeria. In the case of Algeria in fact, 74% of the total studied content is in dialectal Arabic. This rate corresponds to 58% regarding the Tunisian study. Table 6 also shows the strong use of the Latin script for transcribing the dialectal Arabic, with rates reaching more than 81% of the Tunisian dialect productions on the social web (and more than 62% of the studied Algerian dialect productions) that are transcribed in the Latin alphabet. Several factors have been mentioned in (Younes et al. 2015) that can explain this trend of using the Latin alphabet for dialectal writing, such as the scarcity of Arabic keyboards in the early years of the web and mobile devices, multilingualism, influence of colonization, migration, and neo-cultures.

Samih and Maier (2016a) for their part, reported that the construction of an Arabic Moroccan code-switched corpus from textual productions, transcribed in Arabic and extracted from Moroccan discussion forums and blogs, resulted in a corpus comprising 35% of Darija words (Moroccan Arabic) and 49% of MSA words. The rest of the words making up the corpus consists of words from other languages (French, English, Spanish or Berber), some ambiguous words, named entities, numbers and punctuations.

It can therefore, be seen that the various Maghrebi dialects are widely used on the social web, in both the Latin and the Arabic scripts. Their automatic processing is gradually attracting a growing interest among the NLP community. MAD NLP however, faces many difficulties and there are still challenges ahead to be met.

2.3 MAD NLP: difficulties and challenges

In dealing with the Maghrebi dialects automatic processing, several difficulties and challenges are to be overcome which are mainly due to:

- The lack of spelling conventions and orthographic standards. For example, the word meaning "*he does not want*" has the following potential Romanised writings: "*mayehibbish*", "*maye7ibbish*", "*ma yhibbish*", "*ma ye7ebbech*", etc. and several arabic writings such as": "مَيْحبش" , "ما يحبش",etc. More examples are shown in Tables 1 and 22.
- Language continual evolution with the borrowing of a variety of new words. Examples are shown in Table 1.
- MAD variety. The Maghrebi dialects vary from one country to another and present dissimilarities on different linguistic levels. They also present significant dissimilarities with MSA, which make it difficult to directly use available MSA NLP tools that would undoubtedly yield to lower performance.
- Language Ambiguity. Diglossia, multilingualism and code-switching characterizing the Maghrebi dialects arise the issue of MAD identification which becomes an intricate task, facing the problem of language ambiguity. Ambiguous words

Table 3 Verb conjugation specificities of MAD

MAD specificity	Example
Use of *n-* ن as the first-person singular prefix on verbs	نَـمْشِي/namchi (*I go*)
Use of n-ن plus suffix و-u for the plural	نَـمْشِيـــو/namchiw (*we go*)
Negation: use of pre-verbal and post-verbal affixes: **ma** verb-**ich**	ما كتبتـش/**ma** ktibt**ich** (*I did not write*)
	ما نكتبـش/**ma** nikt**ibch** (*I don't write*)
Future tense: use of a pre-verbal marker plus the imperfective form of the verb	**pre-verbal**+namchi/نَمْشِي (*I will go*)
Passive form: obtained generally by prefixing the verb with **t**	قتل/qtel "kill"→تـقتل/**t**eqtel (*was killed*)

can be both MSA and MAD words when they are transcribed in Arabic. When they are Romanized, they may share a common writing with French, English and dialectal words. Some examples of ambiguous words are presented in Table 7.

As we can see, MAD NLP raises important issues to be dealt with and its development requires the availability of large and appropriate language resources. This is why, we propose in the reminder of this paper a comprehensive review of the different kinds of constructed MAD LRs, from the study of existing work carried out on these dialects and make a census of the LRs currently available online for their NLP.

3 MAD raw corpora

3.1 Speech corpora

Maghrebi dialect speech processing has piqued interest of several researchers in the last few years, who resorted to oral recordings to construct new corpora. Some of them resorted to recording the speech of different speakers (Pellegrino and Barkat 1999; Barkat 1999; Barkat and Vasilescu 2001; Barkat et al. 2003, 2004; Bezoui et al. 2019) and real conversations between people (Bougrine et al. 2016; Hassine et al. 2016, 2018). Djellab et al. (2017) used records, in particular, including digital radio calls, phone interception systems, voicemails, etc. Others collected extracted speech from TV shows and Radios (Bougrine et al. 2015; Lachachi and Adla 2015, 2016a, b; Amazouz et al. 2017), from broadcast news (Ali et al. 2016) web streamed local radio channels and TVs and Youtube channels (Bougrine et al. 2017). In (Bougrine et al. 2016), a speech corpus for Algerian Arabic subdialects (named "*ALG-DARIDJAH*") was collected using spontaneous speech, translated MSA speech and image narration. Table 8 shows the constructed MAD speech corpora.

Table 4 Examples of MAD distinctive linguistic specificities (Pereira 2005, 2011; Sayahi 2014)

Linguistic specificities	MD	AD	TD	LD	HD
At phonological level					
Pronunciation of ذ	د	د	ذ	د	ظ
Pronunciation of ث	ت	ت	ث	ت	ذ
At morphological level					
Prefix for future tense	غَ, غَاد, غَادِ	رايح, رايجين	بِش, باش, ماش	بِ	لَهِ
Example: "I will go"	غَـنمْشي	رايَحنَمْشي	باشِنِمْشي	بِـ نِمْشي	لَهِـنمْشي
	Ghanemchi	**rayah** namchi	**bech** nimchi	**bi**nimchi	**Lahi**nimchi
Distinctive vocabulary	zouin-زوين (*beautiful*)	chbeb-شباب (*beautiful*)	meziane-مزيان (*beautiful*)	kways-كويس (good)	zayn زَيْن (*beautiful*)
	bezzaf- بزاف(*plenty*)	bezzaf- بزاف(*plenty*)	barcha- برشة (*plenty*)	halba-هَلْبَ (*plenty*)	yassir ياسر (*plenty*)
	bhit-بحيت (*because*)	walou-ولو (*nothing*)	najjem- نجّم (*I can*)	gaamiz-قَعمِز (to sit)	hawn هَوْن (*here*)
	wakha-واخا (*ok*)	gaa- قاع (*all*)	bahi-باهي (*good*)	amta-أَمْتى (when)	eheh أهه (*yes*)
	daba- دابا (*now*)	mliha مليحة (good)	fisaa فيسع (*fast*)	bess-بَسْ (however)	ebdei أبْد (*no*)
	safi-صافِ (*enough*)	nahdar نَهْدر (speak)	aslama عالسلامة (*hello*)	zay-زَي (as)	ehnâti أحْناتِ (*we*)
	ghadda-غَدْا (*tomorrow*)	nsaksi نسَقْسي(ask)	bislama بالسلامة (*good bye*)	miya-مِيْة (hundred)	entumâti أنْتُماتِ (*you*)

Table 5 Percentage of Facebook users by language used (Mourtada and Salem 2014; Salem 2017)

	Mauritania (%)		Morocco (%)		Algeria (%)		Tunisia (%)		Libya (%)		Average (%)	
	2014	2016	2014	2016	2014	2016	2014	2016	2014	2016	2014	2016
Arabic	48	61.5	33	70.4	32	74.6	18	68.6	85	93.3	43.2	73.7
English	11	13.6	13	15.3	11	11.04	15	19.5	22	16.8	14.4	15.2
French	59	58.9	75	68.4	76	68.2	91	93.1	2	2.03	60.6	58.2

Table 6 Use of dialects and their transcriptions in the social web: cases of Algeria and Tunisia

Content extracted from the social web[a]	Dialectal content out of total content (%)	Latin script out of dialectal content (%)	Arabic script out of dialectal content (%)
Algeria	74	62.1	37.9
Tunisia	58	81.3	18.7

[a] The content extracted from the social web and from which, these figures are calculated, consists of: The vocabulary's words corresponding to a corpus of 1.1 M of YouTube video comments, regarding the Algerian study (Abidi and Smaili 2017).
A corpus of 66,098 messages, mainly extracted from different Tunisian Facebook pages (politics, media, sport, etc.), regarding the Tunisian study (Younes et al. 2015)

3.2 Speech transcription

Transcribing oral conversations was one of the first approaches that have been followed to overcome the lack of written MAD LRs. This was the case of Iskra et al. (2004) who participated to OrienTel project which aimed to develop speech databases and phonetic standards across Northern Africa, the Middle East and the Arabian Gulf. Two speech corpora were thus created for TD and MD which were subsequently transcribed. A similar work was carried out by Belgacem (2009) who participated in the project "Oréodule: a system board real-time recognition, translation and speech synthesis Arabic" in which he established Arabic multi-dialect corpora including Moroccan, Algerian and Tunisian with other Arabic dialects. He used Transcriber tool for the transcription process. Other speech corpora transcription works has been done and mainly concerned the Tunisian dialect. These works include those of (Graja et al. 2010; Masmoudi et al. 2014a, b; Boujelbane et al. 2015). For Algerian, we cite the work of Meftouh et al. (2012) and Amazouz et al. (2017, b, 2018a, b).

Wray and Ali (2015) studied the crowdsourcing for the speech transcription of dialectal Arabic, including Maghrebi. They collected dialectal speech of debate and news programs uploaded from Aljazeera website and resorted to CrowdFlower (CF) to group the speech utterances according to the dialect. The speech was transcribed automatically using QATS (Ali et al. 2014) and via CF by introducing quality control parameters. Table 9 summarizes the identified transcribed corpora.

Table 7 Examples of ambiguous MAD words

Word	خاطر		*Bard*		*Flous*	
Meaning	**TD**	**MSA**	**TD**	**English**	**TD**	**French**
	Because	*spirit*	*cold*	*Poet*	*money*	*Fuzzy*
Word	تقشير		***Bark***		***Pila***	
Meaning	**AD**	**MSA**	**AD**	**English**	**AD**	**French**
	socks	*peeling*	*only*	*Woof*	*battery*	***crush***
Word	صافي		***Dial***		***Barba***	
Meaning	**MD**	**MSA**	**MD**	**English**	**MD**	**French**
	enough	*clean*	*of*	*Call*	*Beet*	*bother*

3.3 Web and social media corpora

Given the increasing use of Internet in the Arab world and the proliferation of various user-generated content in the Arabic language and its dialects on the web, several researchers have been rather resorting to web and social media to construct various Arabic and Maghrebi dialect LRs.

Several corpora have been thus, extracted from the web and the social web in particular. Some of them cover Arabic dialects in general, including some Maghrebi dialects (Callan et al. 2009; Almeman and Lee 2013; Suwaileh et al. 2016). Others relate exclusively to the MADs, covering only one particular dialect such as the Tunisian dialect (McNeil and Faiza 2011; McNeil 2015; Younes and Souissi 2014; Younes et al. 2015; Bouchlaghem et al. 2014; Masmoudi et al. 2017; Torjmen and Haddar 2018a, b), the Algerian dialect (Abidi and Menacer 2017; Abidi and Smaili 2017, Guellil et al. 2018a, b, c; Soumeur et al. 2018) and the Libyan dialect (Alhammi and Alfards 2018), or covering a subset of MADs (Adouane et al. 2016a). In (Guellil et al. 2018a, b, c) two raw corpora were built from Algerian Facebook pages. The corpora include respectively 8,673,285 messages, 3,668,575 of which are written in Arabic letters and 15,407,910 messages, 7,926,504 of which are written in Arabic and 3,976,700 of which are written in Arabizi.

Characteristics of these LRs as well as their creation context and use are summarized in Table 10.

4 MAD annotated corpora

4.1 Language identification

The identification of the Maghrebi dialects on a token level was the main interest of several researchers, since it is a crucial step for LR construction, especially from the web and social media. In this context several annotated corpora have been generated, where tags mainly correspond to the used language for each considered

unit. Regarding resources dedicated to the identification of a specific MAD, we mention those of (Tratz et al. 2014; Voss et al. 2014; Samih and Maier 2016a, b) and (Tachicart et al. 2017) for Moroccan Arabic, those of (Cotterell et al. 2014; Guellil and Azouaou 2016a; Adouane and Dobnik 2017) and (Lichouri et al. 2018) for Algerian Arabic and that of (Aridhi et al. 2017) for Tunisian Arabic. As for corpora built by (Salama et al. 2014; Cotterell and Callison-Burch 2014; Mubarak and Darwish 2014; Zaidan and Callison-Burch 2014; Sadat et al. 2014a, b; Harrat et al. 2015; Adouane et al. 2016a; Alshutayri and Atwell 2017, 2018a, b; Saadane et al. 2017, 2018; Zaghouani and Charfi 2018; Alsarsour et al. 2018; El-Haj et al. 2018; Alshutayri and Atwell 2018a, b) and (Altamimi et al. 2018), they cover several Arabic dialects including some MADs.

When the identification task is performed at a sentence level, the works usually concern more than one dialect. The annotation type involves the level of dialect that the sentence includes and the topic or the gender. Regarding the word-level identification, the works rather deal with code switching textual contents and the annotations concern the type of dialect and other information like punctuation marks.

Various corpora dedicated to the MAD identification are described in Table 11.

4.2 Morphosyntactic annotation

The morphosyntactic analysis of a language is a crucial step towards the study of its word formation and structure. With regard to Maghrebi dialects, the study of the literature shows that several works have been carried out, on their morphosyntactic analysis. However, we can see that, most of this work focused on the Tunisian dialect in particular including word segmentation (McNeil 2012), sentence segmentation (Zribi et al. 2016), morphological analysis (Graja et al. 2013; Zribi et al. 2013a, 2016, 2017), POS-tagging (Boujelbane et al. 2014; Hamdi et al. 2014; Zribi et al. 2017) and syntactic analysis (Mekki et al. 2017). Other works focused on the morpho-grammatical analysis of Algerian dialect (Harrat et al. 2014, 2016; Guellil and Azouaou 2016b), while Almeman and Lee (2012), Al-Shargi et al. (2016), Eldesouki et al. (2017), Samih et al. (2017) and Darwish et al. (2018a) have considered various Arabic dialects including some Maghrebi dialects. Two works were dedicated to diacritics restoration for Algerian dialect (Harrat et al. 2013) and Moroccan and Tunisian dialect (Darwish et al. 2018b).

Most of the researchers who worked on the morphosyntactic annotation of MAD resorted to MSA existing resources and tools, such as Zribi et al. (2013a), Boujelbane et al. (2014), Harrat et al. (2014). The works were based on the existing resources by adding some changes on the suffixes, prefixes, morpheme's orders and creating new rules for the studied Maghrebi dialect.

Others like Guellil and Azouaou (2016b) resorted to social media to extract the textual content and observed the characteristics of the Algerian dialect in order to syntactically analyse it.

These works resulted in the construction of a number of annotated resources that are described in the following Table 12. The column "Annotation" indicates which level of analysis is used for annotation.

Table 8 MAD speech corpora

Dialect	Authors	Year	Corpus	Source	Context—purpose
MAD+OAD[a]	Pellegrino and Barkat	1999	80 speech samples of duration about 1 min (8 ArD)	Recorded speech (20 speakers, 10 from the Maghreb) a free translation of a short text ('La Bise et le Soleil')	Automatic identification of Arabic dialects
MAD+OAD	Barkat Barkat and Vasilescu Barkat et al.	1999 2001 2003	96 speech samples of varying duration ranging from 7 to 30 s (6 ArD)	Recorded speech (12 speakers, 6 from the Maghreb) short narration by describing, in dialect, a book made of 15 pictures	Automatic identification of Arabic dialects
MAD+OAD	Barkat et al.	2004	750 utterances (10 sentences/subject) with an average duration of 2.5 s (6 ArD)	Recorded speech	Automatic identification of Arabic dialects
AD	Bougrine et al.	2015	*SAADID*: 941 sentences	Recorded speech+speech from TV and radios (64 speakers)	Spoken Algerian dialect identification
AD+MD+TD	Lachachi and Adla	2015, 2016a, b	MD: 54.19 h; TD: 53.37 h; AD-Oran: 49.73 h; AD-Alg: 51.32 h; AD-Const: 45.18 h	Speech from TV (525 speakers)	Spoken Dialect identification
MD+TD	Hassine et al.	2016	40 pronunciations of the digits (0…9)	Recorded speech (4 Tunisian +4 Moroccan speakers)	Maghrebi dialect speech recognition
AD	Bougrine et al.	2016	*ALG-DARIDJAH*: 4h30 of speech; 6213 utterances	Recorded speech (109 speakers)	Building Speech Corpus and spoken Dialect Identification

Table 8 continued

Dialect	Authors	Year	Corpus	Source	Context—purpose
MAD+OAD	Ali et al.	2016	MAD: 9 h—1934 utterances for training and 2 h—335 utterances for test (4 ArD+MSA+English)	Broadcast news+Aljazeera's video server	Spoken Dialect identification
AD+MD+TD	Amazouz et al.	2017	MCSM: 53 h of speech (AD 14 h, MD 15 h, TD 24 h)	Television media containing entertainment and TV talk shows	Analyse Arabic-French code-switching
AD	Djellab et al.	2017	*AMCASC*: 88 h of speech	Digital radio calls, phone interception systems, voicemails, etc.	Automatic regional accent recognition
AD	Bougrine et al.	2017	*KALAM'DZ*: 104.4 h of speech	Web streamed local radio channels and TVs, Youtube channels (4881 speakers)	Building Speech Corpora
TD	Hassine et al.	2018	4 corpora of pronunciations of the digits (0…9)	Recorded speech (10 Tunisian speakers)	Tunisian Dialect Recognition
AD+MD+TD+OAD	Terbeh et al.	2018	Two corpora 413 min 388 min	Recorded speech	Spoken Dialect identification
MD	Bezoui et al.	2019	Corpus of pronunciations of 4 MD words	Recorded speech (20 speakers)	Moroccan Dialect Recognition

[a] MAD is indicated when the authors did not specify the type of dialect

Table 9 MAD Transcribed Corpora

Dialect	Authors	Year	Script	Corpus	Source	Context—purpose
MD+TD+OAD	Iskra et al.	2004	A	*OrienTel* (5 ArD+OL)	Recorded speech (750 Tunisian speakers, 750 Moroccan speakers)	Development of speech databases and phonetic standards
AD+MD+TD+OAD	Belgacem	2009	A	Corpus of 10 h of speech of 8 ArD 37% of which are transcribed MD: 90mn,7% of which are transcribed TD: 90mn 5% of which are transcribed AD: 90mn, 6.7% of which are transcribed (8 ArD)	Recorded speech (radio and television); manual transcription	Real time system for recognition, translation and synthesis of Arabic speech
TD	Graja et al.	2010	A	*TuDiCoI 2010*: 127 dialogues; 893 utterances; 3403 words.	Recorded speech (conversations in railway station); manual transcription	Lexical study of TD Semantic labeling
AD	Meftouh et al.	2012	A	10 h, 30% of which are transcribed	Recorded speech; manual transcription	Machine translation AD-MSA
TD	Masmoudi et al.	2014a, b	A	*TARIC*: 20 h of transcribed speech; 71,684 words	Recorded speech; manual transcription	Speech recognition or a speech synthesis system
TD	Boujelbane et al.	2015	A	Corpus: 5 h of transcribed speech; 12,149 words	Tunisian TV channel; Tunisian constitution	Automatic processing of oral spoken Tunisian
MAD+OAD	Wray and Ali	2015	A	850 h with approximately 18% ArD speech (4 ArD+MSA)	Aljazeera's video server; automatic transcription	Crowd sourcing for speech transcription of dialectal Arabic

Table 9 continued

Dialect	Authors	Year	Script	Corpus	Source	Context—purpose
AD	Amazouz et al.	2017 2018	L	FACTS: 7 h 30 of speech	Recorded speech (13 speakers); manual transcription	Analyse Arabic-French code-switching

Table 10 Web and social media MAD corpora

Dialect	Authors	Year	Script	Corpus	Source	Context—purpose
MAD+OAD	Callan et al.	2009	A	*ArClueWeb09*: 30 M Web pages (0.4 M MAD pages (1.6%)) (4 ArD+MSA+OL)	Web sites	Construct web datasets available for research
TD	McNeil and Faiza McNeil	2011, 2015	A	*TAC* 2011: 400 K words *TAC* 2015: 820 K words	Traditional written sources, Blog, Email, Facebook, Transcribed Audio	Project of a TD-English Dictionary creation
MAD+OAD	Almeman and Lee	2013	A	5 M sentences (1 M MAD sentences 10.1 M MAD tokens) (4 ArD)	Blogs, comments, forums, Facebook, Twitter,	Build multi dialect Arabic texts corpora from web
TD	Bouchlaghem et al.	2014	AL	*MultiTD* Corpus: 32,848 words	Social networks (Twitter, Facebook, etc.), written pieces of theater, dictionaries, transcriptions of spontaneous speech, etc.	Built Tunisian dialect Wordnet
TD	Younes and Souissi	2014	L	Romanized: 43,222 messages	SMS, chat, forum, Facebook comments	Construction of TD LRs
TD	Younes et al.	2015	AL	Romanized: 31,158 messages Arabic: 7145 messages	SMS, chat, forum, Facebook comments	Construction of TD LRs
AD+MD+ TD+OAD	Adouane et al.	2016b	A	Corpus of 230 K words for each dialect (7 ArD +MSA+Berber)	Forums, blogs, Facebook and online newspapers	Identification of dialectal Arabic
MAD+OAD	Suwaileh et al.	2016	A	*ArabicWeb16*: 150 M Web pages MAD: 5 M pages (3%) (4 ArD+MSA+OL)	Wikipedia, Popular Arabic websites, ArClueWeb09, Twitter, etc.	Ad hoc search, question answering, filtering, cross-dialect search, dialect detection, entity search, blog search, and spam detection.

Table 10 continued

Dialect	Authors	Year	Script	Corpus	Source	Context—purpose
AD	Abidi et al.	2017	AL	*CALYOU*: 853 K comments 12.7 M words	Youtube	Analysis on the use of AD in Youtube.
AD	Abidi and Smaili	2017	AL	*CALYOU* enrichment: 1.1 M comments 17.7 M words	Youtube	Analysis on the use of AD in Youtube
TD	Masmoudi et al.	2017	A	Corpus: 100 K words	21,917 words from Blog+8057 words from converted Arabizi corpus 8057 words+Corpus TARIC	Speech recognition system
LD	Alhammi and Alfards	2018	A	5 K tweets	Twitter	Construction of LD LRs
AD	Guellil et al.	2018a, b, c	AL	Two corpus: 15,407,910 messages 8,673,285 messages	Facebook	Transliteration and sentiment analysis
AD	Soumeur et al.	2018	AL	100,000 messages	Facebook	Sentiment analysis
TD	Torjmen and Haddar	2018a, b	A	Test corpus: 18,134 words	Social networks and Tunisian novels.	Morphological Analyzer

Table 11 MAD identification corpora

Dialect	Authors	Year	Script	Corpus	Source	Annotation
MD	Tratz et al.	2014	AL	AMD: 3000 MD tweets RMD: Corpus of (Voss et al. 2014)	Twitter	Word-level Arabic (***Egy/Lev/Gul***) Latin (***MD/Fr/En***)
MD	Voss et al.	2014	L	3300 annotated tweets (63,327 tokens)	Twitter	Word-level (***MD/Fr/En***)
MAD+OAD	Zaidan and Callison-Burch	2014	A	MAD: 0.01% of 3.1M annotated sentences with other Iraqi dialect annotated: 110 K (5 ArD+MSA)	Social media	Sentence-level (level of dialectal content and its type (***Egy, Lev, Gul, Irq, Mag, MSA,…***))
AD	Cotterell et al.	2014	L	339,504 comments; 6,718,502 tokens	Newspaper website	Word-level (***Arabic, Fr, Others***)
MAD+OAD	Salama et al.	2014	A	*YOUDACC*: 640 K annotated sentences (32,215 MAD sentences- 553,900 MAD words) (5 ArD)	User comments on YouTubeS videos	Sentence-level (***Egy, Lev, Gul, Irq, Mag***)
MAD+OAD	Cotterell and Callison-Burch	2014	A	67,468 annotated sentences (6940 annotated MAD comments and 99 annotated MAD tweets) (5 ArD+Others)	Newspapers, twitter	Sentence-level (***Egy, Lev, Gul, Irq, Mag/Fr/others)***
MAD+OAD	Mubarak and Darwish	2014	A	6.5 M annotated tweets (520 K annotated MAD tweets (8%)) (5 ArD+Others)	Twitter	Sentence-level (***ArD)***

Table 11 continued

Dialect	Authors	Year	Script	Corpus	Source	Annotation
AD+HD+LD+MD+TD+OAD	Sadat et al.	2014a, b	A	61,859 sentences AD: 731 sentences; 10,378 words; HD: 2793 sentences; 62,694 words; LD: 370 sentences; 5300 words; MD: 2335 sentences; 30,107 words; TD: 3843 sentences; 18,199 words (18 ArD)	Blogs and forums	Sentence-level (***AD, HD, LD, MD, TD,*** …)
AD+TD+OAD	Harrat et al.	2015	A	Corpus obtained by mixing all the sentences of *PADIC* and indicating for each its language (5 ArD+MSA)	PADIC[a]	Sentence-level (***ALG, ANB, TUN, SYR, PAL, MSA)***
AD	Guellil and Azouaou	2016a	A	100 annotated messages from 18,303 comments	Facebook	Word-level (***AD/NonAD***)
MD	Samih and Maier	2016a, b	A	Corpus: 223 K words 76,732 annotated MD words	Blogs, forums and internet discussions	Word-level (***MSA/Darija/FreSpaBer/***…)
MAD+OL	Adouane et al.	2016a	L	Corpus: 144,535 words (1 ArD+5 OL)	Micro-blogs, forums, blogs and online newspapers	Document-level ***(RB/RA/EN/FR/***…)
TD	Aridhi et al.	2017	L	86,940 annotated words	SMS, chat, forum, Facebook comments	Word-level (***TD/NonTD***)

Table 11 continued

Dialect	Authors	Year	Script	Corpus	Source	Annotation
MD	Tachicart et al.	2017	A	100 K sentences where MSA represents 66% and MD represents 34% MD: 34 K sentences; 370 K tokens	Facebook, TV series, theatrical written plays and transcribed recorded conversations	Word-level (***MD/MSA***)
AD+MD+TD+OAD	Saadane et al.	2017	AL	AD: 78 K messages; MD: 118.7 K messages; TD: 19.1 K messages (4 ArD+MSA)	Online newspapers and Facebook	Word-level (***AD, MD, TD, Egy, MSA***)
AD	Adouane and Dobnik	2017	A	Corpus: 215,843 words for training and 10,107 for test 118,942 AD words	Social media	Word-level (***AD/MSA/Fr/En/Berber/***…)
MAD+OAD	Alshutayri and Atwell	2017	A	210,915 K tweets (41,042 K MAD tweets) 8091 annotated tweets (1585 MAD annotated tweets) (5 ArD)	Twitter	Sentence -level (***Egy, Lev, Gul, Irq, Mag***)
MAD+OAD	Alsarsour et al.	2018	A	*DART*: 144,596 tweets (MAD: 16,350 tweets) 25 K of which are annotated (5 ArD)	Twitter	Sentence-level (***Egy, Lev, Gul, Irq, Mag***)
MAD+OAD	El-Haj et al.	2018	A	16,494 sentences (355,069 words) MAD: 3693 sentences (53,204 word) (4 ArD+MSA)	Arabic Commentary Dataset (AOC) of Zaidan and Callison-Burch (2014)+ TAC corpus of McNeil (2012)	Sentence-level (level of dialectal content and its type (***Egy, Lev, Gul, Mag, MSA***))

Table 11 continued

Dialect	Authors	Year	Script	Corpus	Source	Annotation
AD+LD+MD+TD+OAD	Zaghouani and Charfi	2018	A	*Arap-Tweet*:2.4 M tweets (16 ArD)	Twitter	Sentence-level (Gender, Age, Dialect)
AD+MD+TD+OAD	Saadane et al.	2018	AL	AD: 326 K comments; MD: 120 K comments; TD: 102 K comments (4 ArD+MSA)	Online newspapers and social media	Word-level (***ArD/MSA/***…)
MAD+OAD	Alshutayri and Atwell	2018a, b	A	2290 K documents (41,042 K MAD tweets) 24 K annotated documents (2 K MAD annotated documents) (5 ArD+MSA)	Twitter, Facebook and newspaper web site	Sentence -level Four labels (level of dialectal content, its type (***Egy, Lev, Gul, Irq, Mag***), reason and words**)**
MAD+OAD	Altamimi et al.	2018	A	*BTAC*: 200 K tweets (122 K labelled) (5 ArD+MSA+CA)	Twitter	Sentence-level (dialect, gender, authorship, and topic)
AD	Lichouri et al.	2018	AL	100 sentences translated to 8 Algerian dialects	PADIC	Word and Sentence-level (***ALG, CST, TNS, JLF, KAB, ANB, BTN, DFL***)

[a] Cf. Section 4.3

4.3 Translation

Several works have been carried out on the language translation and have mainly focused on translating MADs into MSA. Among these works, there are those who tackled the translation into MSA, of various Arabic dialects including some MADs (Bouamor et al. 2014, 2018; Meftouh et al. 2015, 2018; Harrat et al. 2015, 2017a; Mubarak 2018). Others focused on translating MADs in particular (Meftouh et al. 2012; Tachicart and Bouzoubaa 2014a, b; Sadat et al. 2014c). As mentioned above, all of them considered dialects as source languages and MSA as the target language. Parallel corpora resulting from these works are described in Table 13.

4.4 Transliteration and code-switching

Transliteration consists in transforming a word from a writing system to another while preserving its pronunciation. As shown in Sect. 2.2, MADs can be written in both Arabic and Latin scripts especially on social networks, blogs and forums. This has led several researchers to focus on the transliteration task, which can be particularly useful for the process of dialectal LR construction and for many other applications such as machine translation (treatment of proper nouns), information retrieval, etc. Works on transliteration of Maghrebi dialects have mainly allowed the construction of various corpora. They tackled Latin to Arabic script transliteration, using pre-established rules (such as in Saadane et al. 2013; Masmoudi et al. 2015; Guellil et al. 2018a), or weighted finite state transducers (such as in Ben Moussa et al. 2019) or machine learning based methods (such as in Younes et al. 2016; Guellil et al. 2017a), or deep learning based Sequence-to-Sequence approach such as in (Younes et al. 2018).

The majority of the works on the transliteration task is based on the Arabic-Latin correspondences between characters. Some of these correspondences were based on rules which were created following an observation of the written content in both Latin and Arabic scripts (Saadane et al. 2013; Guellil et al. 2018a). The work of Masmoudi et al. (2015) was, however, based on CODA which is an orthographic convention for the Tunisian dialect that follows supposedly the same orthographic rules as MSA with some exceptions and extensions.

We summarize LRs resulting from these works in the following Table 14.

Only few corpora of Arabic code-switching have been created by Cotterell (2014), Amazouz (2017) and Samih and Maier (2016a, b).

4.5 Sentiment analysis

Sentiment analysis (also called opinion mining) is used to determine the emotional tone of a user's language productions. It is mainly used to better grasp the perception and the opinions declared in a user statement. It is extremely effective as it provides an overview of users' opinions about a given topic, especially in social media. With the increasing use of Maghrebi dialects in social media, several works were initiated on MAD sentiment analysis in the last few years. Those who focused on a particular MAD include (Elkhlifi et al. 2014; Zarra et al. 2017; Oussous et al.

Table 12 MAD morphosyntactic analysis corpora

Dialect	Authors	Year	Script	Corpus	Source	Annotation
MAD+OAD	Almeman and Lee	2012	A	2229 different dialects words	Web	Word segmentation (Prefix, Suffix and Stem), type and root
TD	McNeil	2012	L	400 K words 2 K annotated words	TV scripts, forums and traditional folktales	Segment boundaries of word
TD	Graja et al.	2013	A	*TuDiCoI 2013*: 1825 dialogues; 12,182 utterances; 21,682 words. Annotation of 7814 words	TuDiCoI corpus	Lexical normalization, morphological analysis and lemmatization and a synonym
TD	Zribi et al.	2013a	A	*STAC*: 03:20—27,144 words	Radio and TV broadcasts	Sentence boundaries, gender, prefix, suffix, number, person, voice, POS, disfluencies, named entities, etc.
AD	Harrat et al.	2013	A	Vocalized corpus: 4 K pairs of sentences, with 23 K words.	manually written	Diacritics
TD	Boujelbane et al.	2014	A	Test corpora: 2h15 min of transcribed speech—4041 words Training corpora: TD version of the ATB	Arabic Tree Bank corpus	POS tags
TD	Zribi et al.	2015, 2016, 2017	A	*STAC*: 04:50:31—7788 sentences—42,388 words	Radio and TV broadcasts + 30 min of TuDiCoI	Written version: Sentence boundaries, word boundaries, gender, number, person, voice, POS, etc. Speech version: Sentence boundaries and disfluencies
AD	Harrat et al.	2014, 2016	A	Corpora: AD-Alg: 6415 sentences—10,790 words; AD-Anb: 6415 sentences—9688 words AD-MSA: 6400 sentences—38,707 AD words—40,906 MSA words	Recorded conversations, movies and TV shows	Segment boundaries of word, POS, gender, number, person

Table 12 continued

Dialect	Authors	Year	Script	Corpus	Source	Annotation
TD	Hamdi et al.	2015	A	Test corpora: 805 sentences containing 10,746 tokens and 2455 types	PATB, transcribed TV series and political debates, etc.	Lemma and POS tag
AD	Guellil and Azouaou	2016b	L	*ASDA*: N/A	Facebook	POS, Type, etc.
MD+OAD	Al-Shargi et al.	2016	A	MD corpus of 64,170 tokens (2 ArD)	Social media, folktales, transcribed oral interviews and textbooks	Morphological and semantic information
TD	Mekki et al.	2017	A	Corpus: 12 K words; 492 sentences Treebank: 928 syntactic trees	Tunisian constitution	Sentences boundaries, Words Tokenization and syntactic analysis
MAD+OAD	Eldesouki et al. Samih et al.	2017 2017	A	350 tweets for each dialect MAD: 350 tweets—5495 tokens (4 ArD)	Twitter	Word boundaries
MAD+OAD	Darwish et al.	2018a	A	MAD: 350 tweets—5495 tokens (4 ArD)	Corpus of Samih et al. (2017)	Word boundaries+POS Tags
MD+TD	Darwish et al.	2018b	A	8200 verses each dialect (134,324 words for MD and 131,923 words for TD)	New Testament	Fully diacritized

Table 13 Translated MAD corpora

Dialect	Authors	Year	Script	Corpus	Source	Translation
AD	Meftouh et al.	2012	A	Parallel corpus AD-MSA	Transcriptions of recorded speech	MSA
TD	Sadat et al.	2014c	A	Corpus: 50 TD phrases translated into MSA	Tunisian forums and blogs	MSA
TD+OAD	Bouamor et al.	2014	A	Corpus *MPCA*: 2 K sentences for each dialect—10,896 TD words (4483 unique words) (5 ArD+MSA+English)	2000 Egyptian sentences from Egyptian-English corpus of Zbib et al. (2012)	MSA-English
MD	Tachicart and Bouzoubaa	2014a	A	N/A	Television productions scenarios	MSA
AD+TD+OAD	Meftouh et al. Harrat et al.	2015 2017a	A	*PADIC*: 6400 sentences for each dialect and MSA (5 ArD+MSA)	Recorded speech and Algerian TV shows	MSA
AD+MD+TD +OAD	Meftouh et al.	2018	A	*PADIC*: 6400 sentences for each dialect and MSA (6 ArD+MSA)	*PADIC* 2015	MSA
MAD+OAD	Mubarak	2018	A	*Dial2MSA* corpus: 5 K MAD tweets (4 ArD+MSA)	Twitter	MSA
AD+LD+MD +TD+OAD	Bouamor et al.	2018	A	*MADAR* Corpus: CORPUS-25: 2000 sentences for each 25 ArD CORPUS-5: 10,000 sentences for each 5 ArD	Basic Traveling Expression Corpus BTEC of Takezawa et al. (2007)	MSA-French–English

Table 14 Latin to Arabic transliterated MAD corpora

Dialect	Authors	Year	Corpus	Source
MAD+OAD	Saadane et al.	2013	–	Web pages
TD	Masmoudi et al.	2015	70,861 messages—870,904 words, 530 of which are annotated messages	Facebook, SMS and YouTube Comments
TD	Younes et al.	2016	19,763 words	Facebook, SMS, etc.
AD	Guellil et al.	2017a	6233 sentences	From PADIC Corpus
AD	Guellil et al.	2018a	300 messages (2005 words)	Facebook
			50 messages (527 words)	Corpus of (Cotterell et al.2014)
TD	Younes et al.	2018	45,629 words	Facebook, SMS, etc.
TD	Ben Moussa et al.	2019	1000 words	Forums, blogs, Facebook, etc.

2018) that focused on the Moroccan dialect (Mataoui et al. 2016; Rahab et al. 2017, 2018; Guellil et al. 2017b, 2018b; Soumeur et al. 2018) that dealt with the Algerian dialect and (Sayadi et al. 2016; Ameur et al. 2016; Mdhaffar et al. 2017) that worked on the Tunisian dialect. As for Al-kabi et al. (2016), they considered several Arabic dialects including MADs, knowing that Maghrebi dialectal content represents only 0.3% of their corpus.

Table 15 shows MAD corpora dedicated to sentiment analysis. The column "Annotation" indicates the tags (positive, negative, …) used to annotate each unit (word or sentence).

4.6 Several annotations

A few multi-annotated corpora have been created. The corpus of Diab et al. (2010) which deals with Moroccan and other Arabic dialects comprises morphosyntactic analysis and identification annotations. The corpus of Baly et al. (2017) deals with three MAD dialects and includes identification and sentiment analysis annotations. The recent Algerian corpus of Abainia (2019) includes several annotations: code-switching, transliteration, translation, dialects and sub-dialect identification, gender identification, sentiment analysis, abuse detection, named entity recognition, etc.

Baly et al. (2017) started by creating from Twitter a corpus called MD-ArSenTD (Multi-Dialect Arabic Sentiment Twitter Dataset) which represents a multi-dialectal dataset composed of tweets collected from 12 Arab countries including Algeria, Morocco and Tunisia. These tweets are annotated by dialect and sentiment labels that incorporate both polarity and intensity information on a 5-point scale (very negative, negative, neutral, positive and very positive). Baly et al. (2017) used only Egyptian and Emirati tweets to discover distinctive features that could facilitate the sentiment analysis task. They also carried out a comparative evaluation of different sentiment models (SVM and LSTM deep learning model) on Egyptian and Emirati tweets. Table 16 shows the corresponding constructed LRs.

5 Lexica, dictionaries and ontologies

As part of works carried out on the automatic processing of Maghrebi dialects, a set of data LRs have been constructed including various lexica, dictionaries and ontologies. Such resources represent indeed essential material for language studies and NLP tool development, especially when dealing with the translation task. In this LRs' category, most of the constructed MAD resources are bilingual lexica involving MADs and MSA (Meftouh et al. 2012; Boujelbane et al. 2013a, b, 2015; Hamdi et al. 2014; Tachicart et al. 2014a, b; Elmarakshy and Ismail 2015; El Abdouli et al. 2019) or MADs and other languages such as English (Graff and Maamouri 2012) and French (Guellil and Azouaou 2017; Azouaou and Guellil 2017) or bilingual lexicon of transliterations (Younes et al. 2015; Abidi and Smaili 2018). Multilingual lexicons were also built covering MADs and other languages (Saadane et al. 2018; Bouamor et al. 2018). The main monolingual MAD lexica mainly include a Tunisian dialect phonetic dictionary (TunDPDic), constructed by Masmoudi et al. (2014c), lexicon for Al-Khalil analyser of (Boudlal et al. 2011) constructed by Zribi et al. (2013b), lexicon for morphological analyser of Torjmen and Haddar (2018a, b), Lexical Database Adaptation for Comp-Dial system realized by Neifar et al. (2014) and Tunisian Arabic lexical dictionary by Ben Moussa et al. (2016). Lexicons for sentiment analysis were also built by (Mataoui et al. 2016; Ameur et al. 2016; Guellil et al. 2017b; Guellil et al. 2018b, c).

As for ontologies involving MADs, they are very scarce and are still at a preliminary stage. In reviewing the literature, we mainly identified domain ontologies related to the Tunisian dialect. These include domain ontologies built as part of works on the processing of TD in dialogue systems (in Railway stations) (Graja et al. 2011a, b; Karoui et al. 2013a, b; Graja et al. 2015). They also include the Tunisian dialect Wordnet "TunDiaWN" proposed by Bouchlaghem et al. (2014) who preserved the AWN content (Arabic Wordnet by Elkateb et al. (2006)), and the "aebWordnet" proposed by Ben Moussa et al. (2014, 2015), Ben Moussa and Alimi (2016), which was modeled from a bilingual dictionary (Tunisian-English). Mrini and Bond (2017) built the Moroccan Darija Wordnet (MDW) using a bilingual Moroccan-English dictionary of (Harrell 1963). These various MAD resources are described in Table 17.

6 Codification and normalization

The informal nature of the dialects makes their automatic processing a challenging task. It is, in fact, a language that doesn't conform to linguistic or orthographic rules and encompasses everyday new terms. Therefore, several researchers resorted to an intermediate step that may ease the processing of such languages, namely the orthographic conventions. This was the idea of Zribi et al.'s work (2013b), they proposed OTTA, an orthographical convention for the transcription of the spoken Tunisian Arabic. They resorted to MSA orthographic rules and defined new ones for the TD specificities on the corpus TuDiCoI (Graja et al. 2010). Another convention

Table 15 Annotated MAD corpora for sentiment analysis

Dialect	Authors	Year	Script	Corpus	Source	Annotation
MD	ElKhlifi et al.	2014	A	*MDOES*: 340 K tweets+5 K comments from FB→5 K annotated sentences	Twitter, Facebook, political discourse	Sentence-level (*Positive, Negative*)
MAD+OAD	Al-Kabi et al.	2016	AL	1442 posts—63,115 words (5 ArD+MSA+OL)	Web site "Maktoob Yahoo!"	Sentence-level (*Positive, Negative, Neutral, Irregular, Unknown)*
TD	Sayadi et al.	2016	A	TEC: 5514 tweets (49,940 words—10,553 unique words), 1754 of which are tweets with polarity	Twitter	Sentence-level (P*ositive, Negative, Neutral)*
AD	Mataoui et al.	2016	AL	Corpus: 7698 comments	Facebook	Sentence-level *(Positive, Negative)*
TD	Ameur et al.	2016	AL	Test corpus: 755 words	Facebook	Word-level (*Surprised, Satisfied, Happy, Gleeful, Romantic, Disappointed, Sad, Angry, Disgusted)*
MD	Zarra et al.	2017	AL	34,576 posts	Facebook	Sentence-level (*Positive, Negative)*
TD	Mdhaffar et al.	2017	AL	*TSAC*: 17 K comments	Facebook	Word and Sentence-level (*Positive, Negative)*
AD	Rahab et al.	2017	A	*ARAACOM*: N/A	Online journals	Sentence-level (*Positive, Negative)*
AD	Guellil et al.	2017b	A	*323 messages* *426 messages*	PADIC Facebook	Sentence-level (*Positive, Negative)*
AD	Guellil et al.	2018b	AL	*8000 messages*	Facebook	Sentence-level *(Positive, Negative)*

Table 15 continued

Dialect	Authors	Year	Script	Corpus	Source	Annotation
AD	Soumeur et al.	2018	AL	25,475 messages	Facebook	Sentence-level (*Positive, Negative, Neutral)*
MD	Oussous et al.	2018	A	MSAC: 2 K comments	Press web site, Facebook, Twitter, and YouTube	Sentence-level (*Positive, Negative)*
AD	Rahab et al.	2019	A	SANA: N/A	Comments of 3 Algerian Arabic newspaper web sites	Sentence-level (*Positive, Negative)*

Table 16 Several annotated MAD corpora

Dialect	Authors	Year	Script	Corpus	Source	Annotation
MD+OAD	Diab et al.	2010	A	*COLABA*: N/A (4 ArD)	Blog	Sentence boundary, for each word degree of dialectness, lemma, PosTag, etc.
AD+MD+TD+OAD	Baly et al.	2017	A	*MD-ArSenTD*: 14,400 tweets (1200 for each dialects) (12 ArD)	Twitter	Sentence-level Identification (*dialect's country and region*) Sentiment analysis *(very negative, negative, neutral, positive, very positive)*
AD	Abainia	2019	L	DZDC12: 2.4 k texts (44.4 k words)	Social media	Code-switching, transliteration, translation, dialects and sub-dialect identification, gender identification, sentiment analysis, abuse detection, named entity recognition, etc.

was developed by Zribi et al. (2014), inspired from the CODA project (a Conventional Orthography for the Dialectal Arabic) (Habash et al. 2012) to create a CODA, specific to the Tunisian dialect. The CODA goals were summarized by Zribi et al. in five perspectives: (1) every word has a single orthographic interpretation; (2) created for computational purposes; (3) uses the Arabic script; (4) intended as a framework for writing all the Arabic dialects; (5) aims to create a balance between establishing conventions based on the MSA-ArD similarities and maintaining a level of dialectal uniqueness. The TD follows supposedly the same orthographic rules as MSA with some phonological, phono-lexical, morphological and lexical exceptions and extensions. Saadane and Habash (2015) and Turki et al. (2016) adapted the same convention and developed respectively an Algerian Arabic CODA and a Maghrebi Arabic CODA. Boujelbane et al. (2016) proposed an automatic process for spontaneously spelled Tunisian Arabic (TA) normalization into the conventional orthography CODA (Zribi et al. 2014). This process is baptized COTA orthography (COnventionalized Tunisian Arabic orthography). Their proposed approach is close to the approach of (Eskander and Habash 2013). They showed that the rule-based and the statistical methods can reduce the transcription errors. Habash et al. (2018) presented a unified set of guidelines and resources for conventional orthography of dialectal Arabic applied to 28 Arab city dialects including 7 Maghrebi cities (Rabat, Fes, Algiers, Tunis, Sfax, Tripoli and Benghazi). The resources of this new CODA*[9] are all available online.

We provide in Table 18, a recap of the various codification and normalization conventions used for MADs.

7 Online available MAD LRs

By examining language resource management platforms such as those of the Language Data Consortium (LDC) and the European Language Resources Association (ELRA), we could see that very few corpora are currently available for Maghrebi dialects, when compared with other languages and also, when compared with MSA, Levantine and Egyptian dialects. In fact, simple queries[10] for macro-Arabic[11] resources available in the LDC catalog show 164 resources. Only 14 LRs involve the MADs. With the ELRA search engine, we found 108 resources. Only 6 involve the MADs. Table 19 shows the details.

Regarding the various identified LRs that have been generated by existing work on MADs we examined for this survey, it is worth noting, that only 23% of total resources are currently available online. We list these available resources in Table 20 and indicate whether they are freely downloadable.

[9] The project website: http://resources.camel-lab.com/.

[10] Queries realized on February 14, 2018.

[11] Return resources in all associated individual languages in addition to any resource in the base macrolanguage.

Table 17 MAD Lexica, dictionaries and ontologies

Type	Dialect	Authors	Year	Script	LR	Source
Lexica and dictionaries	MD+OAD	Graff and Maamouri	2012	AL	Bilingual lexicon MD-Eng MD: 3014 consonantal skeleton classes (3 ArD+English)	A Dictionary of Moroccan-English, English-Moroccan (Harrell and Bergman 2004)
	AD	Meftouh et al.	2012	A	Bilingual lexicon MSA-AD	Transcriptions of recorded speech
	TD	Boujelbane et al.	2013a, b, 2015	A	Bilingual lexicon MSA-TD (1500 verb lemmas, 1050 noun lemmas)	Arabic Tree Bank (ATB) (Maamouri et al. 2004)
	TD	Hamdi et al.	2013a, b	A	Lexicon of 1638 TD-MSA roots-patterns Lexicon of 1500 TD-MSA forms	Arabic Tree Bank (ATB) (Maamouri et al. 2004)
	TD	Sadat et al.	2014c	A	Lexicon: 1600 TD-MSA entries	Tunisian forums and blogs
	TD	Hamdi et al.	2014	A	Bilingual lexicon MSA-TD of deverbal nouns (137,199 nominal entries	TD-MSA lexicon (Boujelbane et al. 2013a)
	MD	Tachicart et al.	2014b	A	Bilingual lexicon MD-MSA *MDED*: 18 K entries	MSA electronic dictionary Almaknaz +physical copy of a MD dictionary Colin+ Web
	MAD+OAD	Elmarakshy and Ismail	2015	AL	Bilingual lexicon MAD-MSA *DALex*: MAD: 1116 entries (4 ArD)	Blog posts and forum articles, the literature works and religious stories
	AD	Guellil and Azouaou	2017	L	Bilingual lexicon AD-French	Social media
		Azouaou and Guellil	2017		25,086 entries	
	AD+MD+ TD+OAD	Saadane et al.	2018	AL	Multilingual Lexicon (4 ArD+MSA) AD: 58,237 entries; MD: 42,282 entries; TD: 43,690 entries	Online newspapers and social media

Table 17 continued

Type	Dialect	Authors	Year	Script	LR	Source
	AD+LD+MD +TD+OAD	Bouamor et al.	2018	A	*MADAR* Lexicon Multilingual: 1045 entries (25 ArD+MSA+French+English)	Basic traveling expression corpus BTEC of Takezawa et al. (2007)
	MD	El Abdouli et al.	2019	AL	Bilingual dictionary MD-MSA	N/A
	TD	Zribi et al.	2013b	A	TD Lexicon for Al-Khalil analyser	
	TD	Masmoudi et al.	2014c	A	Phonetic lexicon *TunDPDic*: 18 K words	TARIC
	TD	Neifar et al.	2014	A	Lexical database adaptation for Comp-Dial system: 1066 of vocabulary, 190 compound words and 182 Verbs	TuDiCoI
	AD	Harrat et al.	2014, 2016	A	Dictionaries for BAMA-MA of Buckwalter (2002)	Recorded conversations, movies and TV shows
	TD	Hamdi et al.	2015	A	Lexicon of verbs: 1638 entries Lexicon of deverbal nouns: 33,271 entries Lexicon of particles: 200 pairs (MSA, TD)	PATB, transcribed TV series and political debates, etc.
	TD	Younes et al.	2015	AL	Lexicon Latin→Arabic 19,763 entries Lexicon Arabic→Latin 18,153 entries	SMS+Chat+Forum+Facebook
	MAD	Adouane et al.	2016a	L	Lexicon of 42 k words with spellings	Micro-blogs, forums, blogs and online newspapers
	AD	Mataoui et al.	2016	AL	Lexicon for sentiment analysis: 2380 positive words—713 negative words	Facebook
	TD	Ameur et al.	2016	AL	Lexicon for sentiment analysis: 131,937 words	Facebook
	TD	Ben Moussa et al.	2016	A	Tunisian Arabic lexical dictionary	Labeled grammatical morphemes
	AD	Guellil et al.	2017b	A	Lexicon for sentiment analysis: - SOCALALG contains 2375 annotated terms (from -5 to +5) - SentiALG contains 3408 annotated terms (from −5 to +5)	English Lexicons: SentiWordNet[a] and SOCAL[b]

Table 17 continued

Type	Dialect	Authors	Year	Script	LR	Source
	AD	Guellil et al.	2018b, c	AL	Lexicon for sentiment analysis: contains 4873 annotated terms (from −5 to +5)	English Lexicon: SOCAL
	TD	Torjmen and Haddar	2018a, b	A	Dictionary for morphological analyser (29 interrogative adverbs, 21 adverbs other types, 9 demonstrative pronouns and more than 1000 verbs)	Extracted from a corpus and a set of morphological local grammars
	AD	Abidi and Smaili	2018	AL	*ADL* Lexicon of 6947 entries with transliterations	Youtube
Ontologies	TD	Graja et al.	2011a, b	A	15 concepts	TuDiCoI
	TD	Karoui et al.	2013a, b	A	*RIO*: 14 concepts, 25 relations, 387 instances	TuDiCoI
	TD	Bouchlaghem et al.	2014	AL	WordNet *TunDiaWN*: 32,848 words entity types: Synset, word, form	MultiTD Corpus
	TD	Ben Moussa et al.	2014	AL	*aebWordnet*: 8279 lemmas	Bilingual English–Tunisian Arabic dictionary «Peace corps dictionary»
		Ben Moussa and Alimi	2015, 2016		Synset: 18,209 entries (8279 lemmas and 25,748 Word-sense pairs)	
	TD	Graja et al.	2015	A	18 concepts	TuDiCoI
	MD	Mrini and Bond	2017	L	*MDW*: 2540 Moroccan Synsets	Bilingual Moroccan-English dictionary (Harrell 1963)

[a] http://sentiwordnet.isti.cnr.it/

[b] https://github.com/sfu-discourse-lab/SO-CAL

Table 18 Codification and normalization conventions for MADs

Dialect	Authors	Year	Script	Convention
TD	Zribi et al.	2013b	A	TD OTTA
TD	Zribi et al.	2014	A	TD CODA
AD	Saadane and Habash	2015	A	AD CODA
MAD	Turki et al.	2016	A	MAD CODA
TD	Boujelbane et al.	2016	A	TD COTA
AD+LD+MD+TD+OAD	Habash et al.	2018	A	CODA* (28 ArD)

Table 19 Available LRs in the LDC and ELRA platforms

Language	LDC[a]	ELRA[b]
Standard	109	N/A
Egyptian	17	5
Levantine	14	5
Gulf	5	0
Mesopotamian	2	0
Moroccan	3	3
Tunisian (with other dialects)	11	3
Algerian	0	0
Arabic and its dialects	164	108

[a] Only Egyptian, Gulf, Mesopotamian, Moroccan, Levantine and Standard were mentioned in language Field. For Tunisian, we introduced in the query "Tunisian" as key in "Find keywords in corpus description" field, knowing that the result returned was multi-dialect resources

[b] In the ELRA search engine, if a field does not exist for a language, we mention the language as a key

Finally, it should be added that, in order to collect data and construct MAD LRs, some of the works examined in this study, have resorted to freely available resources that we list in Table 21.

8 Discussion

8.1 Work on MAD LRs evolution

It is clear from this study that Maghrebi dialects are giving rise to an increasing interest from several NLP researchers. We should however note that, the total number of works that led to the construction of MAD LRs we could identify till May 2019, is of the order of a hundred (exactly 148 works that have led to the construction of 158 different RLs, 143 of which are written RLs), which actually represents a very limited number in comparison with other languages.

In this discussion, we distinguish between works that focused specifically on a single dialect (*AD*: Algerian dialect, *HD*: dialect of Mauritania, *LD*: Libyan dialect, *MD*: Moroccan dialect, *TD*: Tunisian dialect), works which dealt with a subset of specified (named) Maghrebi dialects (*MAD*) and those which focused on a subset of Maghrebi dialects without specifying their nature (*NS MAD*: non-specified MAD).

As we can see in Fig. 1, the evolution of the number of research works witnessed a growth from 2010 to 2018. This figure shows that the evolution varies from one dialect to another and that, it is on the Tunisian dialect that there has been the most work published during these last nine years. This can also be seen in Fig. 2 that gives the breakdown of published works over the targeted dialects. We can indeed observe that, out of the total number of the identified works involving the construction of MAD LRs, 63% focused specifically on a single dialect, mainly on the TD (33%), the AD (20%) and the MD (9%). On the other hand, only one recent work has resulted in a constructed LD LR (Alhammi and Alfard 2018) (1%) and to the best of our knowledge, no works have been carried out specifically on the HD to this day. HD was however treated once in the work of Sadat et al. (2014a, b) that dealt with a set of MADs including HD. Libyan and Mauritanian dialects are thus, currently non-resourced languages. As for multi Arabic dialectal resources including various MADs, we have identified 55 related works, representing 37% of the total work, knowing that 30 of them (20% of the total work) do not specify the type of the considered MAD.

8.2 MAD LRs Scripts

Despite the significant presence of MAD textual productions using the Latin alphabet especially on the social web, we can see that most of the works presented in this study dealt with Maghrebi dialects written in the Arabic script. Very few of them focused on the Latin script and code-switching, such as Cotterell et al. (2014), Samih and Maier (2016b), Guellil et al. (2017a, 2018b, c), and Younes et al. (2015). This is shown in Fig. 3, where 67% of the related work led to the construction of MAD LRs transcribed exclusively in the Arabic script. Only 23% of these works considered both scripts, while 10% were concerned only with Latin transcribed MAD resources.

One of the main reasons explaining the large focus on the Arabic script, consists of the idea of exploiting its closeness to MSA and resorting accordingly, to existing MSA resources and tools for constructing useful LRs for dialects' processing. This approach has been adopted by several researchers, namely Almeman and Lee (2012), Boujelbane et al. (2013a, b), Hamdi et al. (2013a, b) and Zribi et al. (2013a), etc. While very useful, it should be however highlighted that Arabic transcribed MAD resources are far from being able to cover the needed LRs for the Maghrebi dialects' NLP, especially when dealing with user-generated content on the social web. In Maghrebi countries indeed, and as we have shown in Table 6 (in Sect. 2.2), the Latin script is very used, even more than the Arabic script in some countries such as Tunisia, Algeria and Morocco. The Latin Maghrebi dialect, also referred to as "Romanized" and "Arabizi", is generally vowelized since users include letters such as "a", "e", "o", "i", etc. to designate the Arabic vowels, it is also marked by

Table 20 MAD LRs available online

LR		Year	Dialect	Link	Free
Corpora	*OrienTel* Morocco MCA (Modern Colloquial Arabic) database of Iskra et al.	2004	MD	ELRA: http://catalog.elra.info/	
	OrienTel Tunisia MCA (Modern Colloquial Arabic) database of Iskra et al.	2004	TD	ELRA: http://catalog.elra.info/	
	ArClueWeb09 of Callan et al.	2009	MAD+MSA+OAD	https://lemurproject.org/clueweb09/index.php	
	TuDiCoI corpus of Graja et al.	2010, 2013	TD	https://sites.google.com/site/marwagraja/resources	√
	TAC corpus of McNeil	2012, 2015	TD	http://tunisiya.org/ (for browsing online, not downloadable)	√
	Corpus of Almeman and Lee	2013	MAD+OAD	http://www.cs.bham.ac.uk/kaa846/arabic-multi-dialect-text-corpora.html (non functional)	N/A
	TARIC Corpus of Masmoudi et al.	2014a, b, c	TD	http://www-lium.univ-lemans.fr/bougares/ressources.php https://sites.google.com/site/masmoudiabir/res	√
	STAC Corpus of Zribi et al. (raw and annotated)	2015	TD	https://sites.google.com/site/ineszribi/ressources/corpus	√
	Speech corpus of Wray and Ali	2015	MAD+OAD	http://alt.qcri.org/resources/aljazeeraSpeechCorpus/	√
	PADIC corpus of Meftouh et al.	2015	AD+TD+OAD	https://sourceforge.net/projects/padic/files/	√
	Corpus of Ali et al.	2016	MAD+OAD	http://alt.qcri.org/resources/ArabicDialectIDCorpus/	√
	ArabicWeb16 Suwaileh et al.	2016	MAD+MSA+OAD	http://qufaculty.qu.edu.qa/telsayed/arabicweb16	√

Table 20 continued

LR	Year	Dialect	Link	Free
ALG-DARIDJAH corpus of Bougrine et al.	2016	AD	http://perso.lagh-univ.dz/hcherroun/Alg-Daridja.html	√
TAC corpus de 1500 mots Ben Moussa et al.	2016	TD	https://github.com/nadou12/Intelligent-Tunisian-Arabic-Morphological-Analyzer-evaluation-corpus	√
KALAM'DZ corpus of Bougrine et al.	2017	AD	https://github.com/LIM-MoDos/KalamDZ	√
TSAC corpus of Mdhaffar et al.	2017	TD	https://github.com/fbougares/TSAC	√
Corpus of Tachicart et al.	2017	MD	http://arabic.emi.ac.ma:8080/MCAP	√
Corpus of Samih et al.	2017	MAD+OAD	http://alt.qcri.org/resources/da_resources/	√
Dial2MSA corpus of Mubarak	2018	MAD+OAD	http://alt.qcri.org/~hmubarak/EGY-MGR-LEV-GLF-2-MSA.zip.	√
Corpus of Darwish et al.	2018a	MAD+OAD	https://github.com/qcri/dialectal_arabic_resources	√
MADAR corpus of Bouamor et al.	2018	AD+LD+ MD+ TD+OAD	http://nlp.qatar.cmu.edu/madar/ (for browsing online)	
DART dataset of Alsarsour et al.	2018	MAD+OAD	http://qufaculty.qu.edu.qa/telsayed/datasets/	√
PADIC corpus of Meftouh et al.	2018	AD+MD+ TD+OAD	https://www.researchgate.net/publication/316463706_PADIC_A_Parallel_Arabic_DIalect_Corpus	√
MSAC corpus of Oussous et al.	2018	MD	https://github.com/ososs/Arabic-Sentiment-Analysis-corpus	√
SANA Corpus of Rahab et al.	2019	AD	http://rahab.e-monsite.com/medias/files/corpus.rar	√
DZDC12 corpus of Abainia	2019	AD	https://github.com/xprogramer/DZDC12	√
Corpus of El-Haj et al.	2018	MAD+OAD	http://www.lancaster.ac.uk/staff/elhaj/corpora_files/ ArabicDialectsDataset.zip	√

Table 20 continued

LR		Year	Dialect	Link	Free
Lexica, Dictionaries and ontologies	Lexicon of Graff and Maamouri	2012	MD+OAD	LDC: https://catalog.ldc.upenn.edu/	N/A
	RIO ontology of Karoui et al.	2013a b	TD	https://sites.google.com/site/marwagraja/resources	√
	TD Lexicon of Zribi et al.	2013a	TD	https://sites.google.com/site/ineszribi/ressources/lexique	√
	MDED Dictionary of Tachicart et al.	2014a	MD	http://arabic.emi.ac.ma:8080/mded/ (for browsing online, not downloadable)	√
	aebWordnet Ben Moussa et al.	2015	TD	https://github.com/nadou12/aebWordNet-Lexicon	√
	TA lexical dictionary of Ben Moussa et al.	2016	TD	https://github.com/nadou12/Tunisian-Arabic-Lexical-Dictionary	√
	List of strong dialectal words for each major dialect of Mubarak	2018	MAD+ OAD	http://alt.qcri.org/~hmubarak/EGY-MGR-LEV-GLF-StrongWords.zip	√
	ADL lexicon of Abidi and Smaili	2018	AD	http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/Lexicon (non functional)	
	MADAR lexicon of Bouamor et al.	2018	AD+LD+ MD+TD +OAD	http://nlp.qatar.cmu.edu/madar/ (for browsing online)	

Table 21 Other MAD available LRs

LR	Year	Dialect	Link and description	Used in
Text				
New Tunisian Constitution (Klibi Salsabil, Hamraoui Salwa, Ben Abda Hana, Gaddes Chawki, Horcheni Farhat, Maalla Anouar)	2014	TD	https://www.babnet.net/9/destourderjaaa.pdf 12 K words distributed among 492 sentences.	Mekki et al. 2017
Selected documents in literal Arabic and dialectal Arabic (Mohammad Bakri)	2010	MAD+MSA+OAD	http://www.langue-arabe.fr/spip.php?article25 various areas: songs, theatre, newspaper articles	Only TD by Bouchlaghem et al. 2014
Dictionaries				
Algerian to French Dictionary (Said Dzayri)	2013	AD	https://www.flipsnack.com/95C5C758B7A/dictionnaire-arabe-algerien.html 228 verbs, 73 adjectives, 297 nouns and expressions	Guellil and Azouaou 2016b
Tunisian to French Dictionary	N/A	TD	ArabeTunisien.com Over 4 k words and expressions	Bouchlaghem et al. 2014
Tunisian to French Dictionary "le Karmous" (Karim Abdellatif)	2010	TD	https://www.fichier-pdf.fr/2010/08/31/m14401m/dico-karmous.pdf 3800 TD words and proverbs and expressions	Bouchlaghem et al. 2014

Table 21 continued

LR	Year	Dialect	Link and description	Used in
Mo3jam dictionary (user-generated dictionary of colloquial Arabic produced by Abdullah Arif)	N/A	AD+HD+LD+MD+TD+OAD	ar.mo3jam.com AD: 307 terms, HD: 20 terms, LD: 108 terms MD: 495 terms, TD: 102 terms	Only MD by Samih and Maier 2016a All by Suwaileh et al. 2016
Moroccan Darija Dictionary (Abdeljabbar Taoufikallah)	N/A	MD	darijadictionary.com Over 500 entries	Samih and Maier 2016a
English and Tunisian dictionary: Peace Corps dictionary (Rached Ben Abdelkader, Abdeljelil Ayed and Aziza Naouar)	1977	TD	https://files.eric.ed.gov/fulltext/ED183017.pdf (scanned hand-written manuscript)	Ben Moussa et al. 2015

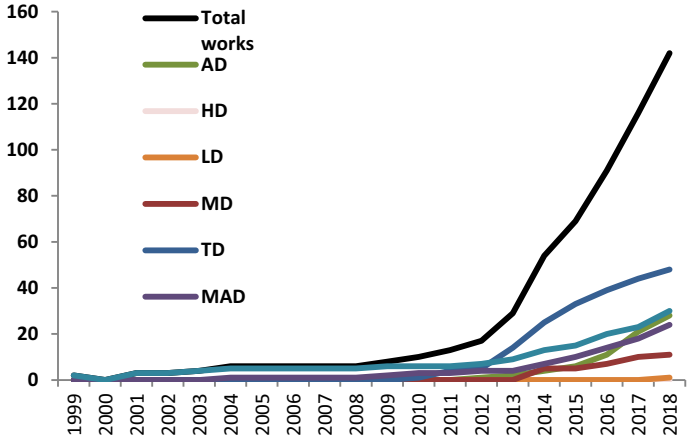

Fig. 1 Evolution of the works on MAD LRs' construction

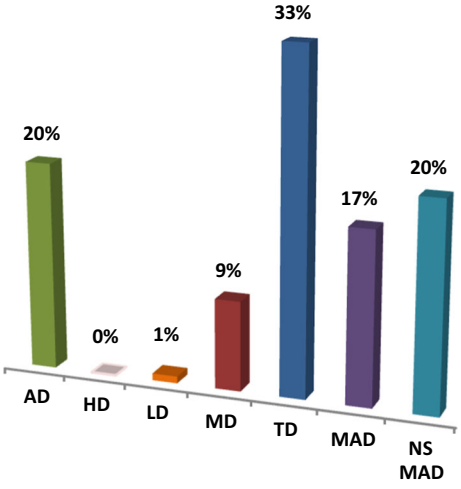

Fig. 2 Distribution of total work according to the target dialect

several other phenomena such as the use of digits to replace Arabic letters with no equivalents in the Latin alphabet, and encompassing abbreviations and acronyms. Therefore, dealing with it would be different, especially in the morphosyntactic and the identification tasks. This form of the dialect still lacks consistent and large corpora allowing its processing, especially regarding the tasks that need the original user's input such as opinion mining. The link between the two forms of MAD can be achieved using the transliteration task. The work on this task can benefit from the Latin script as it includes vowels, unlike the Arabic script. Latin MAD to Arabic

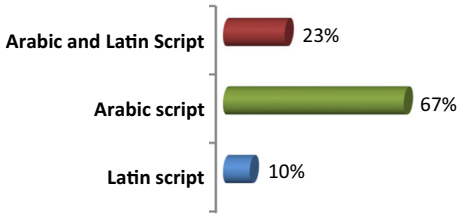

Fig. 3 Distribution of total work according to MADs' script

MAD transliteration was however dealt with in very few works [only in (Saadane et al. 2013; Masmoudi et al. 2015; Younes et al. 2016; Guellil et al. 2017a, 2018a)], while it may constitute a crucial task allowing to automatically generate Arabic MAD LRs from the widely produced Latin content in social media, and helpful for many other kinds of MAD NLP applications (Information retrieval, sentiment analysis, machine translation, etc.). Latin to Arabic transliteration of MADs is to our opinion still an open task and needs to be further explored and studied.

8.3 MAD LRs types

Works studied in this paper have all resulted in the production of various MAD data-LRs including corpora (raw or annotated), lexicons, dictionaries and ontologies. Most of them were built in a given context with the objective of a specific task or application such as translation, segmentation, morphosyntactic analysis, sentiment analysis, etc.

As regards MAD corpora, we identified 36 works that have focused on creating raw corpora which are crucial to study and process Maghrebi dialects. As shown in Fig. 4, 45% of them are speech corpora and 55% of them are written corpora, mainly collected from the web and social media or transcribed from speech.

Based on the study of existing work, it was since 2010 that researchers have been interested in constructing written linguistic resources for Maghrebi dialects. One of the first approaches they have been following, was based on speech transcription, usually performed manually. We can cite among these works (Graja et al. 2010; Masmoudi et al. 2014a), which were mainly carried out on the TD and some subsets of MADs (Fig. 5). Although the transcription approach is a way to address the lack of written LRs, it represents an expensive and time-consuming task, leading to the construction of relatively small corpora. Transcription is also, usually carried out by a native speaker of the considered language or following a transcription convention (Masmoudi et al. 2014a) and, in both cases, the procedure results in a unique transcription of each word. The use of this kind of LRs can therefore be very limited, since they may not be appropriate when dealing with dialectal content produced on social media where, both Latin and Arabic scripts are used, and writing doesn't conform to any spelling rules or conventions. Abidi et al. (2017) illustrate this phenomenon by the example of the word "يرحمك" (that means "*Bless you*") which can be written in 66 different ways with Latin Script. We give in Table 22

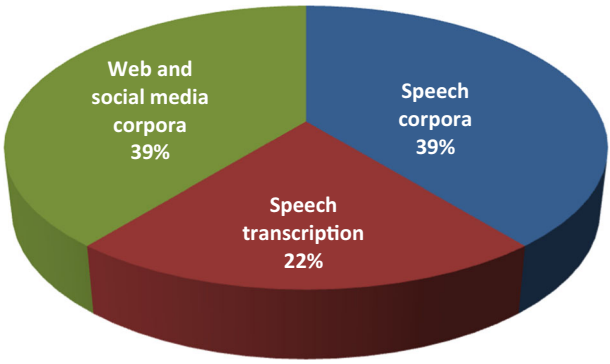

Fig. 4 Distribution of the works on raw corpora by type

other examples of a MAD words written in both Latin and Arabic scripts along with its potential transcriptions.

With the use of social media growth in Maghreb countries and the proliferation of written dialects on the web, several researchers resorted to the web and social media for collecting MAD corpora. Their work represents, according to Fig. 4, 39% of total MAD raw corpora construction works and 64% of total written MAD raw corpora construction works.

Nevertheless, the issues related to intellectual property and legal implications that may arise from the use of social media have not, to our knowledge, been discussed by the various works that have used such data sources. Indeed, the legal issues may arise cross all levels of LR acquisition from their creation to their exploitation:

- During the access to the data, when it comes to collecting the consent of the users;
- During the recordings for speech, where they are related to the respect of the private life, the will of the users and the consequent choice of what it is to show;
- When it comes to anonymizing the content by removing all identity information.
- When it is stored;
- When it is exploited.

Currently, the copyright and ethics of the data collection from social media and the jurisdictions are not always very clear, and they may differ from one country to another. The best appropriate approach would be to have an agreement with the owners of these sites as part of a partnership. Knowing that there are more permissive sites than others, it's easier to work with tweets (using public API). With Facebook, it's a little more complicated even if the extraction API is public as well. We believe that ethical issues could be avoided as long as:

- The data is extracted via a public API. There is no hacking or intrusion into sites to which we are not entitled;
- The data is extracted from public pages and not from closed groups;

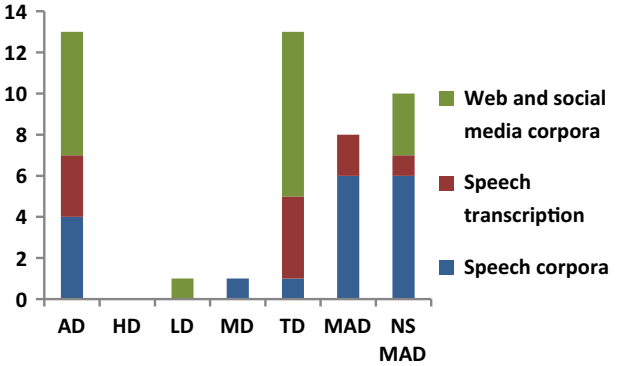

Fig. 5 Distribution of the works on raw corpora according to the target dialect

- The user anonymity is respected and no information concerning his identity is extracted, revealed or kept;
- The extracted data is exclusively intended for research in an academic setting with no commercial purpose;
- The data will not be kept to self by the researcher but will be made available for the research community.

A significant proportion (39%) of the constructed raw corpora are not specific to one particular dialect but include a subset of various MADs. These resources are often constructed as part of works dealing with Arabic dialects in general and including some MADs. Some of them (22% of total work on raw corpora) do not specify the nature (country) of the considered MADs (Mubarak and Darwish 2014; Alshutayri and Atwell 2017, etc.). As for the works which focused on a specific mono-dialectal corpus, they represent 61% of the total work and according to Fig. 5, they concern mainly the TD and AD (thirteen works for each one), followed by the MD and LD (one work). We note that the number of this kind of raw corpora is still very limited and doesn't cover all Maghrebi dialects. More effort should be therefore, put into collecting significant amounts of data for building large corpora. For this, the web and social media seem today to be the richest source of dialectal data that could be utilized in for this purpose. These corpora are crucial to study MAD dialects, allowing both their qualitative and quantitative analysis, building language models and learning their morphology. Moreover, it is from these primary resources that we will be able to produce various annotated resources useful for specific MAD NLP applications.

By examining the works of literature which have resulted in MAD annotated corpora, we identified six types of performed annotations: identification, morphosyntactic analysis, transliteration, translation, sentiment analysis and multi-annotations. Figures 6 and 7 illustrate their distribution.

It can be observed that the majority of the studied works (61%) focused on the identification of dialects and their morpho-grammatical analysis. The problem of

Table 22 Example of a MAD word transcriptions

Writing system	Dialect	Meaning	Transcriptions
Arabic script	AD	Enough	برك | بـرك | بـركة
	MD	Tomorrow	غد | غدا | غدة
	LD	Plenty	هلبا | هالبـة | هالبَ | هلبـة | هلبَ
	TD	Like this	هكا | هكة | هاكة | هاكا
Latin script	AD	All	gā3 | ga | gaa |
	MD	Ok	wakha | wakkha | wa5a | wa55a | wekha | wekkha | we5a |we55a
	LD	To sit	gaamiz | gamiz | ga3miz | gaamez | gamez | ga3mez| gaamèz | gamèz | ga3mèz
	TD	This one	hethi | hedhi | hathi | hadhi | héthi | hèthi | hèdhi | hédhi | hedhy | hédhy | hathy

identification is crucial and arises in the automatic extraction of dialectal content from social media that are highly multilingual. It is mainly the task of distinguishing between MAD dialects and MSA (Samih and Maier 2016a, b; Tachicart et al. 2017; Saadane et al. 2017, 2018) when the considered script is Arabic, and distinguishing between MAD dialects and other languages like French and English when the considered script is Latin (Tratz et al. 2014; Adouane et al. 2016a; Cotterell et al. 2014). As for the morpho-grammatical analysis, it constitutes one of the first essential steps of the linguistic analysis of these dialects and it was the subject of an important part of the works on the MAD (26%) which led to the creation of annotated corpora. We note, however, and as shown in Fig. 7, that these works mainly concerned the Tunisian dialect, followed by the Algerian and Moroccan, while no works of morpho-syntactic analysis were identified for other Maghrebi dialects. We can also note from both Figs. 6 and 7 that 37% of the works leading to annotated corpora mainly concern the following three applications: sentiment analysis, translation and transliteration. As regards the sentiment analysis, the works are relatively recent and led to the construction of annotated corpora, whose numbers and sizes are still relatively small.

All translation works involved the translation of dialects to MSA, often motivated by the idea of applying MSA NLP tools for dialectal processing. As for transliteration, interest in this task is relatively recent (since 2013). Only four works were identified, and they all concerned the Latin to Arabic sense. They mainly focused on the Tunisian and Algerian dialects, given the massive use of the Latin alphabet on social networks in these two countries. Finally, in terms of the number of works that led to annotated corpora, it is the TD that has been the most processed while no annotated corpora were identified for the Libyan and Mauritanian dialects.

Note that only two works concern multi annotated corpora realized by Baly et al. (2017) and Abainia (2019).

The multi annotated corpora presented in this paper are still relatively small. This is due to the difficulty of building such corpora knowing that only the team of

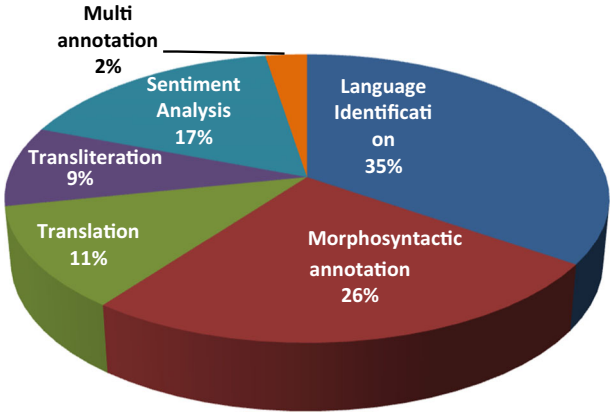

Fig. 6 Distribution of works on annotated corpora by annotation type

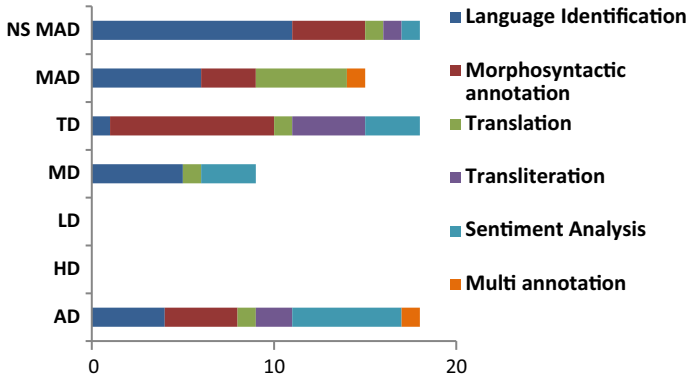

Fig. 7 Distribution of works on annotated corpora by annotation type and dialect

Guellil et al. (2018b, c) proposed a system for automatic construction of annotated sentiment analysis corpora for AD and the team of Alshutayri and Atwell (2018a, b) developed an online game for Arabic dialect annotation of AD identification.

Apart from the raw corpora (oral and written), and annotated corpora, other MAD data resources were constructed as part of various recent works on Maghrebi dialects. These are resources such as lexica, dictionaries and various ontologies. Figures 8 and 9 below illustrate the distribution of these works by type of resource and concerned dialect.

For the ontologies and as shown in Fig. 9, they were built only for TD. These are essentially the domain ontology TudiCol (Graja et al. 2011a, b; Karoui et al. 2013a, b) and the wordnets TunDiaWN (Bouchlaghem et al. 2014) and aebWornet (Ben Moussa et al. 2014, 2016; Ben Moussa et al. 2016). On the other hand, with regard to monolingual and multilingual lexica and dictionaries, we can see that they exist for different dialects like TD, AD, MD and sets of MADs combined, in the number of 33 in all. Efforts were provided by Guellil et al. (2017b, 2018b, c) and Mataoui et al. (2016) for AD and by Ameur et al. (2016) for TD to construct various lexicons for sentiment analysis. Apart from the two bilingual dictionaries MD ↔ EN of Graff and Maamouri (2012) and AD ↔ FR of Guellil and Azouaou (2017), all identified bilingual lexica present essentially a MAD ↔ MSA correspondence. Again, efforts to translate MAD to MSA can be explained by the desire to use the MSA as an intermediate language to apply existing tools of MSA NLP for the processing of MADs or their translation into other languages, like French or English.

Finally, some works focused on the establishment of rules and orthographic conventions. They were mainly carried out for TD (three works), AD (one work) and MAD (one work).

8.4 MAD LRs availability

Regarding the current availability of MAD LRs on the web, we can see from Table 20 of Sect. 7, that despite the efforts made in the various works dealing with

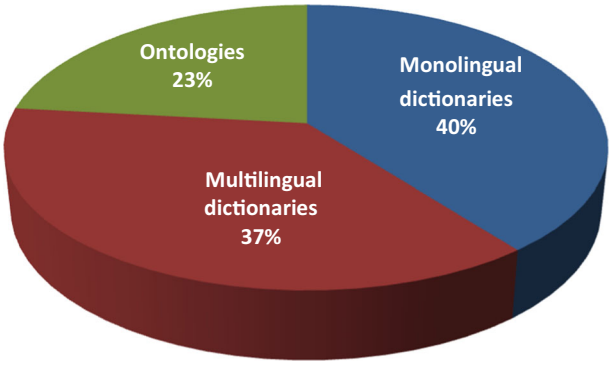

Fig. 8 Distribution of works on lexica, dictionaries and ontologies by type

the processing of MADs, LRs like corpora, lexica, dictionaries, etc., which are available online, are still relatively limited in size and number (23% of the total of resources). Indeed, and if we look for example to the freely available lexica, their size varies between 1 and 18 K words. Annotated corpora have sizes that do not exceed 370 K words for some, 10 K sentences, 17 K Facebook comments and 16 K tweets for others. As for the number of freely available LRs on the web and according to Table 20, we can see that TD is the most available with ten LRs available online, followed by the AD and the MD (both with two LRs available online). Only two multilingual LRs for browsing online and including LD have been identified. For HD, no available RLs have been found.

We can also notice from the studied works, that some constructed resources do not have a wide coverage and are limited to a particular domain. Thus, the corpus built by Graja et al. (2010), Karoui et al. (2013a, b), Neifar et al. (2014) for example, included interactions between staff and clients in Railway stations. The content was limited to the vocabulary of the booking, the tickets, the time and the price. Another

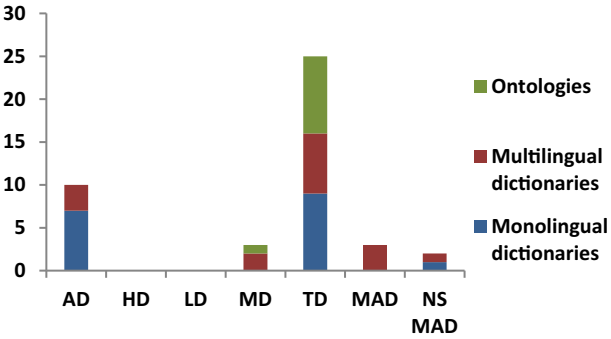

Fig. 9 Distribution of works on lexica, dictionaries and ontologies by type and dialect

Table 23 Examples of MAD words and their origins

MAD Word in AS	MAD Word in LS	Meaning	Etymology	Origin Word
بابور	Babour	Boat	Turkish	Vapur
بابغيو	Babaghayou	Parrot	Italian	Pappagallo
فكرون	Fakroun	Turtle	Berber	Kfrun
ريفز	Rivez	Revise	French	Réviser
تشاتي	tchati	Chatting	English	Chatting
كوردا	courda	Rope	Spanish	Cuerda

work focused on a limited vocabulary (Hassine et al. 2016) and targeted the pronunciation of the digits from 0 to 9. Guellil and Azouaou (2016b) also crawled a unique Facebook page's content namely an Algerian phone operator.

Building resources of significant size and coverage and making them available, remains therefore a priority to allow the study and the processing of the various Maghrebi dialects, especially since the web (the social web in particular) is now an important source of dialectal data.

8.5 Main used approaches in works on MADs

Several approaches were followed for corpus and lexicon creation. For speech corpora, researchers resorted to recording speeches and real conversations between people and collected extracted speech from web streamed local radio channels, TVs and Youtube channels. Regarding textual corpora, some works were based on speech transcription. Others focused on data collection from Social media (Facebook, twitter, User comments on YouTube videos, Blogs and forums, etc.) and online newspaper websites. Given the wide use of social media in Maghreb countries and given the richness of these platforms in dialectal content, they remain the most used source of data collection. They also present an opportunity to build sizable language resources.

To construct lexicons and dictionaries, the majority of researchers started from the corpora in order to cover real and authentic vocabulary.

Regarding annotation, most of the works were performed manually. Only few followed an automatic approach given the difficulty of the tasks and the informality of the MAD language which require the human intervention in all cases.

Maghrebi dialects' automatic processing has been relying on MSA corpora and tools in several works. We can cite Boujelbane et al. (2013a) who exploited an existing MSA lexicon to create TD LRs, Almeman and Lee (2012) who adapted an MSA morphological analyser to deal with the Arabic dialects Harrat et al. (2014, 2016), who built an Algerian dialect dictionary by exploiting the MSA characteristics, etc. Indeed, processing the Maghrebi dialects is not a trivial task as they do not conform to specific syntactic rules or orthographic conventions. Therefore, researches have been resorting to the MSA to exploit its closeness to the dialects, since they originally derive from it. However, several problems may arise following this approach. The main issue is related to the dialectal words' etymologies. In fact, Maghrebi dialects, as mentioned in Sect. 2, are characterized by a linguistic interference between MSA and the neighbouring languages and the influence of colonization (example: French), migration (example: Spanish, Turkish) and neo-cultures (example: Italian). Thus, many words used in Maghrebi dialects are not originally derived from MSA. As we can see through the examples given in Table 23, many user-generated words composing the Maghrebi dialects have no relation with MSA and derive from a completely foreign language.

Moreover, as the phenomenon of multilingualism that the Arab world is witnessing is in a continuous growth, the Maghrebi dialect is undergoing remarkable changes. Indeed, many Maghrebi dialect words possess the roots of foreign languages to which are added new suffixes and prefixes. Some of the foreign words' roots undergo a

change in morphology as well and result on new dialectal words that can be conjugated. For example, the word ("ماتيليشرجيتش"—"matélécharg itech") meaning "I didn't download it", originates from the French verb "télécharger" (to download), to which are added the prefix ("ما"—"ma") and the suffix ("يتش"—"itech"). Therefore, the idea of completely relying on the analogy with MSA is not utterly appropriate since we cannot limit the changes that the Maghrebi dialects undergo with the emergence of new terms of foreign languages. Almeman and Lee (2012) show that the MSA morphology analyser could analyse only 32% of the dialectical words. Building resources and tools specifically dedicated to dialects processing seems thus, to be essential.

Some NLP tasks that were tackled in the studied works, such as dialectal identification, sentiment analysis, labeling and transliteration, used various methods of machine learning (relying notably on NB, SVM, CRF, etc.). On the other hand, we have found that approaches based on deep learning, currently popular in the NLP field, are still very little used and have not yet been explored in the works on the MADs. This could be explained, among others, by the unavailability of the large resources required for the effectiveness of these methods.

9 Conclusion

In this study, we proposed a detailed review of the different kinds of LRs that have been generated as part of various work carried out on the automatic processing of Maghrebi dialects. Our objective was to provide a clear picture of what progress has been made towards constructing LRs for the MADs' NLP and their availability to researchers wishing to work in this field.

From this study, we can conclude that the Maghrebi dialects are receiving increasing interest from the NLP community and that the TD remains the most studied dialect with 33% of the presented works. Indeed, the first works on Maghrebi dialects were carried out on TD (Example: Iskra et al. 2004) which implied the early availability of LRs for this dialect compared to the other MADs that encouraged researchers to further study the TD.

Another important conclusion concerns the writing systems of the constructed LRs. Most of the presented studies were carried out on the Arabic script that was targeted in 67% of the works against 10% for the Latin script and 23% for both Arabic and Latin. The significant focus on the Arabic script is due to the exploitation of available MSA resources given the lack of dialectal content that allows further studies on MADs. We believe that the MADs are still under-resources when it comes to the Latin script and that future works should focus more on Arabizi.

Regarding the used methods for LR construction, we noticed that researchers are becoming more oriented to social media (39% of the MAD raw corpora) since they encompass rich and varied dialectal content.

In terms of availability, only 23% of the works are accessible online which is a low percentage compared to the total works.

Indeed, this survey clearly demonstrated that in recent years, MAD language processing has been generating an increasing interest among NLP researchers.

However, despite these recent efforts, large and freely available MAD NLP dedicated-LRs are still lacking and Maghrebi dialects in general are still among low-resource languages. Some of them, such as the Libyan and Mauritanian dialects are non-resourced languages.

Currently, and given the wide use of social networks in Maghreb countries, the social web seems to be a rich source of dialectal content, that could be utilized to collect data and build large linguistic resources. The efforts to be made in constructing the required resources must, however, take into account the linguistic specificities of these dialects and their informality especially when they are spontaneously produced on the web.

It would be insightful to better consider the written form of these dialects in the Latin script, which is increasingly present on the web. This form generally including vowels, unlike the form transcribed in Arabic, could, thanks to transliteration, contribute significantly to the automatic generation of dialectal Arabic resources and be beneficial for many tasks (such as morphological and syntactic analysis) as well as applications, such as machine translation and sentiment analysis.

It is also essential that the constructed LRs be made available to researchers, in order to be able to make significant progress in the study and the treatment of these dialects, by allowing in particular, to implement various deep learning techniques which so far, have been very little explored in the MADs' NLP.

References

Abainia, K. (2019). DZDC12: A new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-019-09454-8.

Abidi, K., Menacer, M. A., & Smaili. K. (2017). Calyou: A comparable spoken Algerian corpus harvested from youtube. In *Proceedings of the 8th annual conference of the international communication association (Interspeech)*. Stockholm.

Abidi, K., & Smaili, K. (2017). An empirical study of the Algerian dialect of Social network. In *Proceedings of international conference on natural language, signal and speech processing*. Casablanca—Morocco.

Abidi, K., & Smaili, K. (2018). An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Adouane, W., & Dobnik, S. (2017). Identification of Languages in Algerian Arabic Multilingual Documents. In *Proceedings of the third Arabic natural language processing workshop* (pp. 1–8). Valencia, Spain.

Adouane, W., Semmar, N., & Johansson, R. (2016a). Romanized berber and romanized arabic automatic language identification using machine learning. In *Proceedings of the 3rd workshop on NLP for similar languages, varieties and dialects*. Osaka, Japan.

Adouane, W., Semmar, N., Johansson, R., & Bobicev, V. (2016b). Automatic detection of arabicized berber and arabic varieties. In *Proceedings of the third workshop on NLP for similar languages, varieties and dialects* (pp. 63–72). Osaka, Japan.

Alhammi, H. A., & Alfard, R. A. (2018). Building a twitter social media network corpus for libyan dialect. *International Journal of Computer Electrical Engineering, 10,* 1.

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., & Renals, S. (2016). Automatic dialect detection in arabic broadcast speech. In *Proceedings of interspeech-2016* (pp. 2934–2938). San Francisco, US.

Ali, A., Mubarak, H., & Vogel, S. (2014). Advances in dialectal arabic speech recognition: A study using twitter to improve Egyptian ASR. In *Proceedings of the 11th international workshop on spoken language translation (IWSLT 2014)*. Lake Tahoe, USA.

Al-Kabi, M., Al-Ayyoub, M., Alsmadi, I., & Wahsheh, H. (2016). A prototype for a standard arabic sentiment analysis corpus. *The International Arab Journal of Information Technology, 13*(1), 163–169.

Almeman, K., & Lee, M. G. (2012). Towards developing a multi-dialect morphological analyzer for Arabic. In *Proceedings of the 4th international conference on Arabic language processing*. Rabat, Morocco.

Almeman, K., & Lee, M. (2013). Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In *Proceedings of the 1st international conference on communications, signal processing, and their applications*. Sharjah, United Arab Emirates.

Alsarsour, I., Mohamed, E., Suwaileh, R., & Elsayed, T. (2018). DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., & Rambow. O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the 10th language resources and evaluation conference*. Portoroz, Slovenia.

Alshutayri, A., & Atwell, E. (2017). Exploring twitter as a source of an arabic dialect corpus. *International Journal of Computational Linguistics, 8,* 2.

Alshutayri, A., & Atwell, E. (2018a). Arabic dialects annotation using an online game. In *Proceedings of the 2nd international conference on natural language and speech processing*. Algiers, Algeria.

Alshutayri, A., & Atwell, E. (2018b). Creating an Arabic dialect text corpus by exploring twitter, facebook, and online newspapers. In *Proceedings of the 3rd workshop on open-source Arabic corpora and processing tools*. Miyazaki, Japan.

Altamimi, M., Alruwaili, O., & Teahan, W. J. (2018). BTAC: A twitter corpus for Arabic dialect identification. In *Proceedings of the 6th conference on computer-mediated communication (CMC) and social media corpora (CMC-corpora 2018)*. Antwerp, Belgium.

Amazouz, D., Adda-Decker, M., & Lamel, L. (2017). Addressing code-switching in French/Algerian Arabic speech. In *Proceedings of INTERSPEECH 2017*. Stockholm, Sweden.

Amazouz, D., Adda-Decker, M., & Lamel, L. (2018). The French-Algerian code-switching triggered audio corpus (FACST). In Proceedings of 11th international conference on language resources and evaluation LREC 2018 (pp. 1468–1473). Miyazaki, Japan.

Ameur, H., Jamoussi, S., & Ben Hamadou, A. (2016). Exploiting emoticons to generate emotional dictionaries from Facebook pages. *Intelligent Decision Technologies, Springer, 2016,* 39–49.

Aridhi, C., Achour, H., Souissi, E., & Younes, J. (2017). Word-level identification of romanized tunisian dialect. In *Proceedings of the 22nd international conference on natural language & information systems* (pp. 170–175). Liège, Belgium.

Assiri, A., Emam, A., & Aldossari, H. (2015). *Arabic sentiment analysis: A survey. IJACSA, 6,* 12.

Azouaou, F., & Guellil, I. (2017). ALG/FR: A step by step construction of a lexicon between Algerian Dialect and French. In *Proceedings of the 31st Pacific Asia conference on language, information and computation, PACLIC 31*. Cebu, Philippines.

Barkat, M. (1999). Identification if Arabic dialects and experimental determination of distinctive cues. In *Proceedings of the 14th international congress of phonetic sciences*. San Francisco, US.

Barkat, M., Hamdi, R., & Pellegrino, F. (2004). De la caractérisation linguistique à l'identification automatique des dialectes arabes. In *Proceedings of the MIDL Workshop*. Paris, France.

Barkat, M., & Vasilescu, I. (2001). From perceptual designs to linguistic typology and automatic language identification: Overview and perspectives. In *Proceedings of Eurospeech, 7th European conference on speech communication and technology*. Aalborg, Denmark.

Barkat, M., Vasilescu, I., & Pellegrino, F. (2003). Stratégies perceptuelles et identification automatique des langues. *Revue Parole, 25,* 26.

Belgacem, M. (2009). Construction d'un corpus robuste de différents dialectes arabes. *Actes des 8emes Rencontres Jeunes Chercheurs en Parole*, 33.

Ben Moussa, N. K., & Alimi, A. M. (2015). Construction d'un Wordnet standard pour l'Arabe tunisien. In *Proceedings of Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications*. Sousse, Tunisia.

Ben Moussa, N. K., Soussou, H., Alimi A. M. (2016). Intelligent Tunisian Arabic morphological analyzer. In *Proceedings of the 2016 IEEE/ACS 13th international conference of computer systems and applications (AICCSA).* Agadir, Morocco.

Ben Moussa, N. K., Soussou, H., & Alimi, A. M. (2014). Building a standardized Wordnet in the ISO LMF for aeb language. In *Proceedings of the 7th Global Wordnet Conference (GWC 2014), association for computational linguistics* (pp.71—77). Tartu-Estonia.

Ben Moussa, N. K., Soussou, H., Alimi, A. M. (2015). Tunisian Arabic aebWordnet: Current state and future extensions. In *Proceedings of the first international conference on Arabic computational linguistics*. Cairo, Egypt.

Ben Moussa, N. K., Soussou, H., & Alimi, A. (2019). Tunisian arabic chat alphabet transliteration using probabilistic finite state transducers. *The International Arab Journal of Information Technology, 16,* 2.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2013). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56,* 85–100.

Bezoui, M., Beni Hssane A., & Elmoutaouakkil, A. (2019). Speech recognition of moroccan dialect using hidden Markov models. In *Proceedings of international symposium on machine learning and big data analytics for cybersecurity and privacy (MLBDACP)*. Leuven, Belgium.

Bouamor, H., Habash, N., & Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the ninth international conference on language resources and evaluation*. Iceland, May.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Bouchlaghem, R., Elkhlifi, A., & Faiz, R. (2014). Tunisian dialect Wordnet creation and enrichment using web resources an other Wordnets. In *Proceedings of the EMNLP 2014 Workshop on Arabic natural language processing* (pp. 104—113). Doha, Qatar.

Bougrine, S., Cherroun, H., & Ziadi, D. (2015). Prosody-based Spoken Algerian Arabic dialect identification. In *Proceedings of the international conference on natural language and speech processing*. Algiers, Algeria.

Bougrine, S., Cherroun, H., Ziadi, D., Lakhdari, A., & Chorana, A. (2016). Toward a rich Arabic speech parallel corpus for algerian sub-dialects. In *Proceedings of the 2nd workshop on Arabic corpora and processing tools 2016 theme: Social Media*. Portorož, Slovenia.

Bougrine, S., Chorana, A., Lakhdari, A., & Cherroun, H. (2017). Toward a web-based speech corpus for Algerian Arabic dialectal varieties. In *Proceedings of the 3rd Arabic natural language processing workshop (WANLP)* (pp. 138—146). Valencia, Spain.

Boujelbane, R., Khemakhem, M. E., Béchet, F., & Belguith, L. H. (2015). De l'arabe standard vers l'arabe dialectal: Projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens. *Revue TAL, 55,* 2.

Boujelbane, R., Khemekhem, M. E., & Belguith, L. H. (2013b). Mapping rules for building a Tunisian Dialect Lexicon and generating corpora. In *Proceedings of the international joint conference on natural language processing*. Nagoya, Japan.

Boujelbane, R., Khemekhem, M. E., BenAyed, S., & Belguith, L. H. (2013a). Building Bilingual Lexicon to Create Dialect Tunisian Corpora and Adapt Language Model. In *Proceedings of the 2nd workshop on hybrid approaches to translation, ACL 2013*. Sofia, Bulgaria.

Boujelbane, R., Mallek, M., Khemakhem., M. E., & Belguith L. H. (2014). Fine-grained POS Tagging of Spoken Tunisian Dialect Corpora. In *Proceedings of the 19th international conference on application of natural language to information systems* (pp. 59–62). Montpellier, France.

Boujelbane, R., Zribi, I., Kharroubi, S., & Khemakhem, M. E. (2016). An automatic process for Tunisian Arabic orthography normalization. In *Proceedings of the 10th international conference on natural language processing (HrTAL2016)*. Dubrovnik, Croatia.

Callan, J., Hoy, M., Yoo, C., & Zhao, L. (2009). The ClueWeb09 Dataset, 2009. *Presentation Nov. 19, 2009 at NIST TREC. Slides online at* boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf.

Cotterell, R., & Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the 9th international conference on language resources and evaluation*. Reykjavik, Iceland.

Cotterell, R., Renduchintala, A., Saphra, N., & Callison-Burch, C. (2014). An Algerian Arabic-French code-switched corpus. In *Proceedings of the workshop on free/open-source arabic corpora and corpora processing tools workshop programme* (pp. 34). Reykjavik, Iceland.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Dimitrov, M., Dowman, M., et al. (2009). *Developing language processing components with GATE Version 5 (a User Guide)*. Sheffield: The University of Sheffield.

Darwish, K., Abdelali, A., Mubarak, H., Samih, Y., & Attia, M. (2018b). Diacritization of Moroccan and Tunisian Arabic dialects: A CRF approach. In *Proceedings of the 3rd workshop on open-source Arabic corpora and processing tools*. Miyazaki, Japan.

Darwish, K., Mubarak, H., Abdelali, A., Eldesouki, M., Samih, Y., Alharbi, R., Attia, M., Magdy, W., & Kallmeyer, L. (2018a). Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. In *Proceedings of the LREC workshop on semitic language processing* (pp. 66—74). Malta.

Djellab, M., Amrouche, A., Bouridane, A., & Mehallegue, N. (2017). Algerian modern colloquial Arabic speech corpus (AMCASC): Regional accents recognition within complex socio-linguistic environments. *Language Resources and Evaluation, 51*(3), 613–641.

Duong, L. (2017). Natural language processing for resource-poor languages. *Ph.D. thesis, the University of Melbourne*. Melbourne, Australia.

Eldesouki, M., Samih, Y., Abdelali, A., Attia, M., Mubarak, H., Darwish, K., & Kallmeyer, L. (2017). Arabic multi-dialect segmentation: bi-LSTM-CRF vs. SVM. *CoRR*, abs/1708.05891.

El-Haj, M., Kruschwitz, U., & Fox, C. (2014). Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation, 46* (3), 549–580.

EL-Haj, M., Rayson, P., & Aboelezz, M. (2018). Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th edition of the language resources and evaluation conference* (pp. 3622—3627). Miyazaki, Japan.

Elimam, A. (2004). Le maghribi, alias ed-darija, langue consensuelle du Maghreb. *éd. Dar El Gharb*. Alger.

Elimam, A. (2009). Du Punique au Maghribi Trajectoires d'une langue sémito-méditerranéenne. *Synergies Tunisie no 1*, 25–38.

Elimam, A. (2012). *Le maghribi, vernaculaire majoritaire à l'épreuve de la minoration*. Oran: ENSET.

Elkateb, S., Black, B., Vossen, P., Farwell, D., Pease, A., & Fellbaum, C. (2006). Arabic WordNet and the challenges of Arabic. In *Proceedings of the challenge of Arabic for NLP/MT conference* (pp. 15—24). London, UK.

Elkhlifi, A., Bouchlaghem, R., & Rhazi, A. (2014). Opinion extraction in Moroccan Dialect Texts. In *Proceedings of the 5th international conference on arabic language processing*. Oujda, Morocco.

Baly R., El-Khourya, G., Moukalleda, R., Aouna, R., Hajja, H., Shabanb, K. B., & El-Hajj, W. (2017). Comparative evaluation of sentiment analysis methods across Arabic dialects. In *Proceedings of the 3rd international conference on arabic computational linguistics, ACLing 2017*, Dubai. United Arab Emirates.

El Abdouli. A., Hassouni, L., Anoun, H. (2019). A distributed approach for mining Moroccan Hashtags using Twitter Platform. In *Proceedings the 2nd international conference on networking, information systems & security*. Rabat, Morocco.

Elmarakshy, R., & Ismail, M.A. (2015). Compiling a dialectal Arabic lexicon Using Latent Topic models. In *Proceedings of the 1st international conference on advanced intelligent system and informatics (AISI2015)*. Beni Suef, Egypt.

Embarki, M. (2008). Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabica, 5*(6), 583–604.

Eskander, R., & Habash. N. (2013). Automatic correction and extension of morphological annotations. In *Proceedings of the 7th linguistic annotation workshop & interoperability with discourse* (pp.1–10). Sofia, Bulgaria.

Fishman, A. J. (1999). *Handbook of language and ethnic identity*. New York: Oxford University Press.

Graff, D., & Maamouri, M. (2012). Developing LMF-XML bilingual dictionaries for colloquial Arabic dialects. In *Proceedings of the 8th international conference on language resources and evaluation* (pp. 269–274). Istanbul, Turkey.

Graja, M., Jaoua, M., & Belguith, L. H. (2010). Lexical study of a spoken dialogue corpus in Tunisian dialect. In *Proceedings of the international arab conference on information technology (ACIT'2010)*. Benghazi-Libya.

Graja, M., Jaoua, M., & Belguith, L. H. (2011a). Building ontologies to understand spoken Tunisian dialect. *International Journal of Computer Science, Engineering and Applications, 1,* 4.

Graja, M., Jaoua, M., & Belguith, L. H. (2011b). Towards understanding Spoken Tunisian dialect. In *Proceedings of the 18th international conference (ICONIP 2011)*. Shanghai, China

Graja, M., Jaoua, M., & Belguith, L. H. (2013). Discriminative framework for spoken Tunisian dialect understanding. In *Proceedings of the first international conference on statistical language and speech processing (SLSP 2013)*. Tarragona, Spain.

Graja, M., Jaoua, M., & Belguith, L. H. (2015). Statistical framework with knowledge base integration for robust speech understanding of the Tunisian dialect. In *IEEE/ACM transactions on audio, speech, and language processing*, 23(12).

Guellil, I., Adeel, A., Azouaou, F., & Hussain, A. (2018b). Sentialg: Automated corpus annotation for Algerian sentiment analysis. In *Proceedings of the international conference on brain inspired cognitive systems* (pp. 557-567).

Guellil, I., Adeel, A., AZOUAOU, F., Hachani, A. E., & Hussain, A. (2018c). Arabizi sentiment analysis based on transliteration and automatic corpus annotation. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 335-341). Brussels, Belgium.

Guellil, I., & Azouaou, F. (2016a). Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect. In *Proceedings of the 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)* (pp. 724–731).

Guellil, I., & Azouaou, F. (2016b). ASDA: Analyseur Syntaxique du Dialecte Algérien dans un but d'analyse sémantique. In *Proceedings of Conférence Nationale d'Intelligence Artificielle*. Clermont-Ferrand, France.

Guellil, I., & Azouaou, F. (2017). Bilingual Lexicon for Algerian Arabic Dialect Treatment in Social Media. In *Proceedings of WiNLP: Women & underrepresented minorities in natural language processing (co-located with ACL 2017)*. Vancouver, Canada.

Guellil, I., Azouaou, F., Abbas, M., & Sadat, F. (2017a). Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic. In *Proceedings of the first workshop on social media and user generated content machine translation (co-located with EAMT 2017)*. Prague, Czech Republic.

Guellil, I., Azouaou, F., Benali, F., Hachani, A. E., & Saadane, H. (2018a). Approche Hybride pour la translitération de l'arabizi algérien: une étude préliminaire. In Proceedings of the 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN). Rennes, France.

Guellil, I., Azouaou, F., Saâdane, H., & Semmar, N. (2017b). Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien. *La revue internationale Traitement Automatique des Langues (TAL)* (pp. 41–65).

Rahab, H. Zitouni, A., & Djoudi, M. (2017). ARAACOM: ARAbic algerian corpus for opinion mining. In *Proceedings of the 3rd international conference of computing for engineering and sciences*. Istanbul, Turkey.

Habash, N., Diab, M., & Rabmow. O. (2012). Conventional orthography for dialectal Arabic. In *Proceedings of the 8th international conference on language resources and evaluation*. Istanbul, Turkey.

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., & Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the 11th edition of the language resources and evaluation conference. Miyazaki*, Japan.

Hamdi, A., Boujelbane, R., Habash, N., & Nasr, A. (2013a). Un Système de Traduction de Verbes entre Arabe Standard et Arabe Dialectal par Analyse Morphologique Profonde. In *Proceedings of TALN 2013*. Nantes, France.

Hamdi, A., Boujelbane, R., Habash, N., & Nasr, A. (2013b). The Effects of factorizing root and pattern mapping in bidirectional Tunisian—standard Arabic machine translation. In *Proceedings of MT Summit 2013*. Nice, France.

Hamdi, A., Gala, N., & Nasr, A. (2014). Automatically building a Tunisian Lexicon for Deverbal Nouns. In *Proceedings of the first workshop on applying NLP tools to similar languages, Varieties and Dialects* (pp. 95—102). Dublin, Ireland.

Hamdi, A. Nasr, A., Habash, N., & Gala, N. (2015). POS-tagging of Tunisian dialect using standard arabic resources and tools. In *Proceedings of the second workshop on arabic natural language processing* (pp. 59–68). Beijing, China.

Harrat, H., Abbas, M., Meftouh, K., & Smaili, K. (2013). Diacritics restoration for Arabic dialect texts. In *Proceedings of interspeech-2013*. Lyon, France.

Harrat, S., Meftouh, K., & Smaili. K. (2017a). Creating Parallel Arabic dialect corpus: Pitfalls to avoid. In *Proceedings of the 18th international conference on computational linguistics and intelligent text processing (CICLING)*. Budapest, Hungary.

Harrat, S., Meftouh, K., & Smaïli. K. (2017b). Machine translation for Arabic dialects (survey). *Information processing and management*.

Harrat, S., Meftouh, K., & Smaïli. K. (2017c). Maghrebi Arabic dialect processing: An overview. In *Proceedings of the international conference on natural language, signal and speech processing.* Casablanca, Morocco.

Harrat, S., Meftouh, K., Abbas, M., Hidouci, K. W., & Smaili, K. (2016). An algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications, 7,* 3.

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaili, K. (2015). Cross-dialectal Arabic processing. In *Proceedings of the 16th international conference on computational linguistics and intelligent text processing*. Cairo, Egypt.

Harrat, S., Meftouh, K., Abbas, M., & Smaïli, K. (2014). Building resources for algerian arabic dialects. *Corpus (sentences), 4000,* 2415.

Harrell, R. S. (1963). A dictionary of Moroccan Arabic: Moroccan-English. *Georgetown University Press*.

Harrell, R. S., & Bergman, E. M. (2004). A dictionary of Moroccan Arabic: Moroccan-English/English-Moroccan. *Georgetown Classics in Arabic Languages and Linguistics series*.

Hassine, M., Boussaid, L., & Messaoud, H. (2016). Maghrebian dialect recognition based on support vector machines and neural network classifiers. *International Journal of Speech Technology, 19*(4), 987–995.

Hassine, M., Boussaid, L., & Messaoud, H. (2018). Tunisian Dialect Recognition Based on Hybrid Techniques. *The International Arab Journal of Information Technology, 15,* 1.

Iskra, D. J., Siemund, R., Borno, J., Moreno, A., Emam, O., Choukri, K., Gedge, O., Tropf, H., Nogueiras, A., Zitouni, I., Tsopanoglou, A., & Fakotakis, N. (2004). Orientel-telephony databases across northern Africa and the middle east. In *Proceedings of the 4th international conference on language resources and evaluation*. Lisbon, Portugal

Karoui, J., Graja, M., Boudabous, M. M., & Belguith, L. H. (2013a). Domain ontology construction from a Tunisian spoken dialogue corpus. In *Proceedings of the international conference on web and information technologies (ICWIT 2013)*. Hammamet, Tunisia.

Karoui, J., Graja, M., Boudabous, M. M., & Belguith, L. H. (2013b). Semi-automatic domain ontology construction from spoken corpus in Tunisian dialect: Railway request information. *International Journal of Recent Contributions from Engineering, Science & IT, 1*(1), 35–38.

Lachachi, N.-E., & Adla, A. (2015). GMM-Based Maghreb dialect identification system. *Journal of Information Processing Systems., 11*(1), 22–38.

Lachachi N., & Adla A. (2016a). Identification Automatique des Dialectes du Maghreb. *Revue Maghrébine des Langues (RML10)*, 85–101.

Lachachi N., & Adla A. (2016b). Two approaches-based L2-SVMs reduced to MEB problems for dialect identification, *International Journal of Computational Vision and Robotics*.

Lichouri, M., Abbas, M., Freihat, A. A., & Megtouf, D. E. H. (2018). Word-level vs sentence-level language identification: Application to Algerian and arabic dialects. *Procedia Computer Science, 142,* 246–253.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic Corpus. In *Proceedings of NEMLAR conference on Arabic language resources and tools*. Cairo, Egypt.

Masmoudi, A., Bougares, F., Khmekhem, M. E., Estève, Y., & Belguith, L. H. (2017). Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation, 52*(1), 249–267.

Masmoudi, A., Habash, N., Khemakhem, M. E., & Belguith, L. H. (2015). Arabic transliteration of romanized Tunisian dialect text: A preliminary investigation. In *Proceedings of the 16th international conference on intelligent text processing and computational linguistics*. Cairo, Egypt.

Masmoudi, A., Khemakhem, M. E., Estève, Y., Bougares, F., Dabbar, S., & Belguith, L. H. (2014a). Phonétisation automatique du dialecte tunisien. In *Proceedings of JEP 2014*. Le Mans, France.

Masmoudi, A., Khemakhem, M. E., Estève, Y., Belguith, L. H., & Habash, N. (2014b). A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In *Proceedings of the 9th edition of the language resources and evaluation conference*. Reykjavik, Iceland.

Masmoudi, A., Estève, Y., Khmekhem, M. E., Bougares, F., & Belguith, L. H. (2014c). Phonetic Tool for the Tunisian Arabic. In *Proceedings of the 4th international workshop on spoken language technologies for under-resourced languages*. St. Petersburg, Russia.

Mataoui, M., Zelmati, O., & Boumechache, M. (2016). a proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. *Research in Computing Science, 110,* 55–70.

McNeil, K. (2012). Tunisian Arabic Morphological Parser. *Ling-420*.

McNeil, K. (2015). *Tunisian Arabic corpus: A written corpus of an "unwritten" language*. Vienna: International Symposium on Tunisian and Libyan Arabic Dialects, University of Vienna.

McNeil, K., & Faiza, M. (2011). Tunisian Arabic Corpus: Creating a written corpus of an "unwritten" language. In *Proceedings of the Workshop on Arabic Corpus Linguistics*. Lancaster University, UK.

Mdhaffar, S., Bougares, F., Estève, Y., & Belguith, L. H. (2017). Sentiment analysis of Tunisian Dialect: Linguistic Resources and Experiments. In *Proceedings of the 3rd Arabic natural language processing workshop* (pp. 55–61). Valencia, Spain.

Meftouh, K., Bouchemal, N., & Smaïli, K. (2012). A study of a non-resourced language: An Algerian dialect. In *Proceedings of the 3rd international workshop on spoken languages technologies for under-resourced languages* (pp. 125–132). Cape Town, South Africa.

Meftouh, K., Harrat, K., Jamoussi, S., Abbas, M., & Smaili. K. (2015). Machine Translation Experiments on PADIC: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*. Shanghai, China.

Meftouh, K., Harrat, S., & Smaïli, K. (2018). PADIC: Extension and new experiments. In *Proceedings of the 7th international conference on advanced technologies. Antalya, Turkey.*

Mekki, A., Zribi, I., Khemakhem, M. E., & Belguith, L. H. (2017). Syntactic Analysis of the Tunisian Arabic. In *Proceedings of the international workshop on language processing and knowledge management*. Sfax, Tunisia.

Mohand, T. (1999). Substrat et convergences: Le berbère et l'arabe nord-africain. *Estudios de Dialectologi´a Norteaafricana y andalus´ı,* 4, 99–119.

Mourtada, R., & Salem, F. (2014). Citizen engagement and public services in the Arab World: The potential of social media. *Arab Social Media Report series, 6th edition*.

Mrini, K., & Bond. F. (2017). Building the Moroccan darija wordnet (mdw) using bilingual resources. In *Proceedings of the international conference on natural language, signal and speech processing (ICNLSSP)*. Casablanca, Morocco.

Mubarak, H. (2018), Dial2MSA: A Tweets Corpus for Converting Dialectal Arabic to Modern Standard Arabic. In *Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*. Miyazaki, Japan.

Mubarak, H., & Darwish, K. (2014). Using Twitter to Collect a Multi-Dialectal Corpus of Arabic. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing* (pp. 1—7). Doha, Qatar.

Mzoughi, I. (2015). *Intégration des emprunts lexicaux au français en arabe dialectal tunisien*. Linguistique: Université de Cergy Pontoise.

Neifar, W., Bahou, Y., Graja, M., & Jaoua, M. (2014). Implementation of a symbolic method for the Tunisian Dialect understanding. In *Proceedings of the 5th international conference on Arabic language processing (CITALA 2014)*. Oujda, Morocco.

Novotney, S., Schwartz, R., & Khudanpurb, S. (2016). Getting more from automatic transcripts for semi-supervised language modeling. *Computer Speech & Language, 36,* 93–109.

Oussous, A., Lahcen, A. A., & Belfkih, S. (2018). Improving sentiment analysis of Moroccan tweets using ensemble learning. In *Proceedings of the 3rd international conference on big data, cloud and applications* (pp. 91–104). Kenitra, Morocco.

Pellegrino, F., & Barkat, M. (1999). Investigating dialectal differences via vowel system modeling: Application to Arabic. In *Proceedings of the 14th international congress of phonetic sciences*. San Francisco, USA.

Pereira, C. (2005). Arabe maghrébin. In *Proceedings of Actes du Colloque International Langues d'Europe et de la Méditerranée LEM*. Nice, France.

Pereira, C. (2011). Arabic in the North African Region. Stefan Weniger (ed) in collaboration with Geoffrey Khan, Michael P. Streck and Janet C. E. Watson. *Semitic Languages*, 944–959.

Rahab, H., Zitouni, A., & Djoudi, M. (2019). SANA: Sentiment analysis on newspapers comments in Algeria, *Journal of King Saud University—Computer and Information Sciences*, https://doi.org/10.1016/j.jksuci.2019.04.012.

Rosner, M. (2009). Electronic language resources for Maltese. *B. Comrie, R. Fabri, E. Hume, M. Mifsud & M. Vanhove (Eds.), Introducing maltese linguistics. John Benjamins Publishing*, 113, 251-276.

Saadane, H., Guidere, M., & Fluhr, C. (2013). La reconnaissance automatique des dialectes arabes à l'écrit. In *Proceedings of colloque international «Quelle place pour la langue arabe aujourd'hui»* (pp. 18–20).

Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In *Proceedings of the second workshop on ARABIC natural language processing* (pp. 69–79). Beijing, China.

Saadane, H., Nouvel, D., Seffih, H., & Fluhr, C. (2017). Une approche linguistique pour la détection des dialectes arabes. *Actes de TALN 2017*, 2: Articles courts.

Saadane, H.,Seffih, H., Fluhr, C., Choukri, K., & Semmar, N. (2018). Automatic identification of Maghreb Dialects using a dictionary-based approach. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Sadat, F., Kazemi, F., & Farzindar, A. (2014a). Automatic identification of Arabic dialects in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis* (pp. 35–40).

Sadat, F., Kazemi, F., & Farzindar, A. (2014b). Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the second workshop on natural language processing for social media* (pp. 22–27). Dublin, Ireland.

Sadat, F., Mallek, F., Sellami, R., Boudabous, M. M., & Farzindar, A. (2014c). Collaboratively constructed linguistic resources for language variants and their exploitation in NLP applications-the case of Tunisian Arabic and the social media. In *Proceedings of the workshop on lexical and grammatical resources for language processing* (pp. 102). Dublin, Ireland.

Salama, A., Bouamor, H., Mohit, B., & Oflazer, K. (2014). YouDACC: The Youtube dialectal Arabic commentary Corpus. In *Proceedings of the 9th International conference on language resources and evaluation* (pp. 1246—1251). Reykjavik, Iceland.

Salem, F. (2017). Social media and the internet of things towards data-driven policymaking in the Arab world: Potential, limits and concerns. *The Arab Social Media Report, 7,* 462.

Samih, Y., Eldesouki, M., Attia, M., Darwish, K., Abdelali, A., Mubarak, H., & Kallmeyer, L. (2017). Learning from relatives: Unified dialectal Arabic segmentation. In *Proceedings of the 21st conference on computational natural language learning* (pp. 432–441). Vancouver, Canada.

Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L., & Solorio, T. (2016). Multilingual code-switching Identification via LSTM recurrent neural networks. In *Proceedings of the second workshop on computational approaches to code switching* (pp. 50–59). Austin, USA.

Samih, Y., & Maier, W. (2016a). An Arabic-Moroccan Darija Code-Switched Corpus. In *Proceedings of the 10th edition of the language resources and evaluation conference*. Portorož, Slovenia.

Samih, Y., & Maier, W. (2016b). Detecting Code-switching in Moroccan Arabic social media. In *Proceedings of SocialNLP @ IJCAI-2016*. New York, USA.

Sayadi, K., Liwicki, M., Ingold, R., & Bui, M. (2016). Tunisian dialect and modern standard Arabic dataset for sentiment analysis: Tunisian election context. In *Proceedings of the 17th international conference on intelligent text processing and Arabic computational linguistics*. Konya, Turkey.

Sayahi, H. (2014). *Diglossia and language contact: Language variation and change in North Africa*. Cambridge: Cambridge University Press.

Shoufan, A. & Alameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *Proceedings of the second workshop on Arabic natural language processing*. Beijing, China.

Soumeur, A., Mokdadi, M., Guessoum, A., & Daoud, A. (2018). Sentiment analysis of users on social networks: Overcoming the challenge of the loose usages of the Algerian dialect. *Procedia computer science, 142,* 26–37.

Suwaileh, R., Kultlu, M., Fathima, N., Elsayed, T., & Lease, M. (2016). ArabicWeb16: A new crawl for today's Arabic web. In *Proceedings of the 39th annual international ACM SIGIR conference on research and development in information retrieval: SIGIR'16* (pp. 673–676). Pisa, Italy.

Tachicart, R., & Bouzoubaa, K. (2014). A hybrid approach to translate Moroccan Arabic dialect. In *Proceedings of the 9th international conference on intelligent systems, (SITA'14)*. Rabat, Morocco.

Tachicart, R., Bouzoubaa, K., & Jaafar. H. (2014). Building a Moroccan dialect electronic dictionary (MDED). In *Proceedings of the 5th international conference on Arabic language processing (CITALA)*. Oujda, Morocco.

Tachicart, R., Bouzoubaa, K., Lhoussain, A. S., & Jaafar. H. (2017). Automatic identification of Moroccan Colloquial Arabic. In *Proceedings of the 6th international conference on Arabic language processing*. Fez, Morocco.

Takezawa, T., Kikui, G., Mizushima, M., & Sumita, E. (2007). Multilingual spoken language corpus development for communication research. *Computational Linguistics and Chinese Language Processing, 12*(3), 303–324.

Terbeh, N., Maraoui, M., & Zrigui, M. (2018). Arabic dialect identification based on probabilistic-phonetic modeling. *Computación y Sistemas, 22*(3), 863–870.

Torjmen, R., & Haddar, K. (2018a). Morphological analyzer for the Tunisian dialect. In *Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) International workshop on temporal, spatial, and spatio-temporal data mining*, 11107, 180–187.

Torjmen, R., & Haddar, K. (2018b). Construction of morphological grammars for the Tunisian dialect. In *Formalizing natural languages with NooJ 2018 and its natural language processing applications, proceedings of the 12th international conference, NooJ 2018*. Palermo, Italy.

Tratz, S., Briesch, D., Laoudi, J., Voss, C., & Holland, V. M. (2014). Language and dialect identification in social media analysis. *In Proceedings of SPIE sensing technology+applications*. Baltimore, USA.

Turki, H., Adel, I., Daouda, T., & Régragui, N. (2016). A conventional orthography for Maghrebi Arabic. In *Proceedings of the 10th edition of the language resources and evaluation conference*. Portoroz, Slovenia.

Versteegh, K. (1997). *The Arabic language* (p. 277). Columbia: Columbia University Press-Foreign Language Study.

Voss, C., Tratz, S., Laoudiy, J., & Briesch, D. (2014). Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the 9th international conference on language resources and evaluation*. Reykjavik, Iceland

Witt, A., Heid, U., Sasaki, F., & Sérasset, G. (2009). Multilingual language resources and interoperability. *lre-intro.tex; 28/01/2009; 14:31; 2009 Kluwer Academic Publishers*. The Netherlands.

Wray, S., & Ali, A. (2015). Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic. In *Proceedings of Interspeech-2015*. Dresden, Germany.

Younes, J., Achour, H., & Souissi, E. (2015). Constructing linguistic resources for the Tunisian dialect using textual user-generated contents on the social web. In *Proceedings of the 1st international workshop on natural language processing for informal text (NLPIT 2015) in conjunction with the international conference on web engineering (ICWE 2015)*. Rotterdam, The Netherlands.

Younes, J., & Souissi, E. (2014). A quantitative view of Tunisian dialect electronic writing. In *Proceedings of the 5th international conference on Arabic language processing* (pp. 63–72). Oujda, Morocco.

Younes, J., Souissi, E., & Achour, H. (2016). A hidden Markov model for automatic transliteration of romanized Tunisian Dialect. In *Proceedings of the 2nd international conference on arabic computational linguistics*. Konya, Turkey.

Younes, J., Souissi, E., Achour, H., & Ferchichi, A. (2018). A sequence-to-sequence based approach for the double transliteration of Tunisian dialect. *Procedia Computer Science, 142,* 238–245.

Zaghouani, W., & Charfi, A. (2018). Arap-Tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the 11th edition of the language resources and evaluation conference*. Miyazaki, Japan.

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *International Journal of Computational Linguistics (IJCL), 40*(1), 171–202.

Zarra, T., Chiheb, R., Moumen, R., Faizi., R., & ElAfia. A. (2017). Topic and sentiment model applied to the colloquial Arabic: A case study of Maghrebi Arabic. In *Proceedings of the 2017 international conference on smart digital environment* (pp. 174–181). Rabat, Morocco.

Zbib, R., Malchiodi, K., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., & Callison- Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 49–59). Montreal, Canada.

Zribi, I., Boujelbane, R., Masmoudi, A., Khemakhem, M. E., Belguith, L. H., & Habash, N. (2014). A conventional orthography for Tunisian Arabic. In *Proceedings of the 9th edition of the language resources and evaluation conference*. Reykjavik, Iceland.

Zribi, I., Khemakhem, M. E., & Belguith, L. H. (2013a). Morphological analysis of Tunisian Dialect. In *Proceeding of international joint conference on natural language processing (IJCNLP 2013)*. Nagoya, Japan.

Zribi, I., Graja, M., Khemakhem, M. E., Jaoua, M., & Belguith, L. H. (2013b). Orthographic transcription for Spoken Tunisian Arabic. In *Proceedings of the 14th international conference on intelligent text processing and computational linguistics* (pp. 153–163). Samos, Greece.

Zribi, I., Kammoun, I., Khemakhem, M. E., Belguith, L. H. & Blache, P. (2016). Sentence boundary detection for transcribed Tunisian Arabic. In *Proceedings of the 13th conference on natural language processing (KONVENS 2016)*. Varanasi, India

Zribi, I., Khemakhem, M. E., Belguith, L. H., & Blache, P. (2015). Spoken Tunisian Arabic Corpus \STAC: Transcription and annotation. *Research in Computing Science, 90,* 123.

Zribi, I., Khemakhem, M. E., Belguith, L. H., & Blache, P. (2017). Morphological Disambiguation of Tunisian Dialect. *Journal of King Saud University, 29,* 147–155.