The MADAR Arabic Dialect Corpus and Lexicon

Houda Bouamor, Nizar Habash,[†] Mohammad Salameh, Wajdi Zaghouani,[♣]
Owen Rambow,[‡] Dana Abdulrahim,[*] Ossama Obeid,[†] Salam Khalifa,[†]
Fadhl Eryani,[†] Alexander Erdmann,[†] Kemal Oflazer
Carnegie Mellon University in Qatar, Qatar, [♣]Hamad Bin Khalifa University, Qatar
[†]New York University Abu Dhabi, UAE [‡]Columbia University, USA, [*]University of Bahrain, Bahrain
hbouamor@cmu.edu, nizar.habash@nyu.edu, wzaghouano@hbku.edu.qa,
rambow@cs.columbia.edu, darahim@uob.edu.bh

Abstract

In this paper, we present two resources that were created as part of the Multi Arabic Dialect Applications and Resources (MADAR) project. The first is a large parallel corpus of 25 Arabic city dialects in the travel domain. The second is a lexicon of 1,045 concepts with an average of 45 words from 25 cities per concept. These resources are the first of their kind in terms of the breadth of their coverage and the fine location granularity. The focus on cities, as opposed to regions in studying Arabic dialects, opens new avenues to many areas of research from dialectology to dialect identification and machine translation.

Keywords: Arabic Dialects, Parallel Corpus, Lexicon

1. Introduction

Dialectal Arabic (DA) is emerging nowadays as the primary written language of informal communication online in the Arab World: in emails, blogs, discussion forums, chats, SMS, etc. There has been a rising interest in research on computational models of Arabic dialects in the last decade (Meftouh et al., 2015). There have been several efforts on creating different resources to allow building models for several Natural Language Processing (NLP) applications. However, these efforts have been disjoint from each other, and most of them have focused on a small number of dialects that represent vast regions of the Arab World (Zaidan and Callison-Burch, 2011; Diab et al., 2014; Sajjad et al., 2013).

In this paper, we present two resources we created as part of the Multi Arabic Dialect Applications and Resources (MADAR) project.[1] The goal of MADAR is to create, for a large number of dialects, a unified framework with common annotation guidelines and decisions, and targeting applications of Dialect Identification (DID) and Machine Translation (MT).

The first resource is a large parallel corpus of 25 Arabic city dialects, in addition to the pre-existing parallel set for English, French and Modern Standard Arabic (MSA). The second resource is a 25-way lexicon of 1,045 entries in each city's dialect along with MSA, French, and English. These resources are the first of their kind in terms of the breadth of their coverage and their fine granularity. The kind of resources we present in this paper are useful not only for building computational systems but also for studying Arabic dialects from a linguistics perspective (e.g., computational dialectology).

[1]The project site is at http://nlp.qatar.cmu.edu/madar/. مَدَار *madAr* means 'orbit' in Arabic.

2. Arabic and its Dialects

The Arabic language is a family of varieties. Among these varieties, Modern Standard Arabic (MSA) is the shared language of culture, media, and education from Morocco to Oman. However, MSA is not the native language of any speaker of Arabic. In unscripted situations where spoken MSA would typically be required (such as talk shows on TV), speakers usually resort to repeated code-switching between their dialects and MSA (Abu-Melhim, 1991; Bassiouney, 2009). Arabic dialects are often classified regionally (as Egyptian, North African, Levantine, Gulf, Yemeni) (Habash, 2010) or sub-regionally (e.g., Tunisian, Algerian, Lebanese, Syrian, Jordanian, Kuwaiti, Qatari). DA varies phonologically, lexically, and morphologically from MSA, from region to region and to a lesser extent, from city to city in each region (Watson, 2007). In the rest of this section, we discuss these differences.

Phonology An example of phonological differences is in the pronunciation of dialectal words whose MSA cognate has the letter Qaf (ق *q*).[2] It is often observed that in Tunisian Arabic, this consonant appears as /q/ (similar to MSA), while in Egyptian and Levantine Arabic it is /ʔ/ (glottal stop) and in Gulf Arabic it is /g/ (Haeri, 1991; Habash, 2010).

Orthography While MSA has a standard orthography, the dialects do not. Often people write words reflecting the phonology or the etymology of these words. DA is sometimes written in the so-called Arabizi Romanization script (Darwish, 2014). In the context of NLP, a set of conventional orthography guidelines (CODA) has been proposed for a number of dialects (Habash et al., 2012a; Jar-

[2]Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTĎςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ, ŷ ئ, ħ ة, ý ى.

Region	Maghreb				Nile Basin	Levant		Gulf		Yemen
Sub-region	**Morocco**	**Algeria**	**Tunisia**	**Libya**	**Egypt/Sudan**	**South Levant**	**North Levant**	**Iraq**	**Gulf**	**Yemen**
Cities	**Rabat** (*RAB*) Fes (*FES*)	Algiers (*ALG*)	**Tunis** (*TUN*) Sfax (*SFX*)	Tripoli (*TRI*) Benghazi (*BEN*)	**Cairo** (*CAI*) Alexandria (*ALX*) Aswan (*ASW*) Khartoum (*KHA*)	Jerusalem (*JER*) Amman (*AMM*) Salt (*SAL*)	**Beirut** (*BEI*) Damascus (*DAM*) Aleppo (*ALE*)	Mosul (*MOS*) Baghdad (*BAG*) Basra (*BAS*)	**Doha** (*DOH*) Muscat (*MUS*) Riyadh (*RIY*) Jeddah (*JED*)	Sana'a (*SAN*)

Table 1: Different region, sub-region, and city dialects considered in building the MADAR resources.

rar et al., 2014; Zribi et al., 2014; Saadane and Habash, 2015; Turki et al., 2016; Khalifa et al., 2016), and has been recently unified under the CODA* effort (Habash et al., 2018).

Morphology Morphological differences are quite common. One example is the future marker particle which appears as +س *s+* or سوف *swf* in MSA, +ح *H+* or رح *rH* in Levantine dialects, +ه *h+* in Egyptian and باش *bAš* in Tunisian. This together with variation in the templatic morphology make the forms of some verbs rather different: e.g., 'I will write' is سأكتب *sÂktb* (MSA), حأكتب *HÂktub* (Palestinian), هكتب *hktb* (Egyptian) and باش نكتب *bAš nktb* (Tunisian).

Syntax Comparative studies of several Arabic dialects suggest that the syntactic differences between the dialects are relatively minor compared to morphological differences (Brustad, 2000). For example, negation may be realized differently using a combination of prefixes and suffixes (ما *mA*, مش *mish*, مو *muw*, لا *lA*, لم *lam*, etc.) but its syntactic distribution is to a large extent uniform across varieties (Benmamoun, 2012).

Lexicon The number of lexical differences among dialects is significant. The following are a few examples (Habash et al., 2012a): Egyptian بس *bas* 'only' and طربيزة *tarabayzaħ* 'table' correspond to MSA فقط *faqaT* and طاولة *TAwilaħ*, respectively. For comparison, the Levantine forms of the above words are بس *bas* and طاولة *TAwliħ*.

These differences pose serious challenges for Arabic NLP. The challenges are mainly related to the lack of resources and tools. The tools developed for MSA or for a specific dialect cannot effectively model DA which makes its direct use for handling dialects impractical (Habash et al., 2012b).

3. The MADAR Corpus

We built the MADAR Corpus, the first collection of parallel sentences covering the dialects of 25 cities from the Arab World, in addition to English, French, and MSA.[3] Table 1 shows the break up we follow in choosing these cities. This table relates the typical five-way regional break up of Arabic dialects (Habash, 2010) to a more refined ten-way sub-region division, and even further into 25 cities.

The corpus is created by translating selected sentences from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) in French and English to the different dialects.[4] BTEC is a multilingual spoken language corpus containing tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad. This corpus is an attractive resource to use for different reasons: (i) it is conversational in nature (including questions and answers by tourists/guides) and this makes it closer to the genre dialects are used for primarily; (ii) it has short sentences (on the average 6.5 words), which makes it easy enough for the translators to translate; and (iii) the BTEC corpus has translations in several languages which allows the possibility to use this data in the future for training/testing machine translation models across these languages and Arabic dialects.

We selected 2,000 BTEC sentences and translated them to all 25 city dialects (each of these sentences has 25 corresponding parallel translations). Henceforth, we refer to this part of the corpus as CORPUS-25. Furthermore, we selected 10,000 additional sentences and translated them to the dialects of five selected cities: Doha, Beirut, Cairo, Tunis, and Rabat. We call this corpus CORPUS-5. Effectively, each of the five selected cities has 12,000 sentences that are five-way parallel translations, and that could be used to build several Dialectal Arabic NLP applications such as machine translation. An example of a 28-way parallel sentences extracted from CORPUS-25 is given in Figure 1.[5]

Translators, identified from each of the 25 cities specifically, were asked to read a set of sentences provided in English or French, and translate them into their dialects. The translators are all native speakers of the dialects of the cities they hail from. We did not choose MSA as a starting point to avoid biasing the translation (Bouamor et al., 2014). [6]

[3]The MADAR corpus will be made available to the research community. The English part will not be distributed due to copyright restriction. It can be acquired directly from the USTAR consortium (`http://www.ustar-consortium.com/`).

[4]The English, French and MSA versions we use are those provided in the IWSLT evaluation campaign (Eck and Hori, 2005).

[5]The MADAR Corpus is available for browsing online at `http://nlp.qatar.cmu.edu/madar/`.

[6]The translation was handled by Ramitechs (`http://www.ramitechs.com/`), a company that creates and annotates several types of corpora and lexicons using expert linguists.

English	This room is too small.
French	Cette chambre est trop petite.
MSA	هذه الغرفة صغيرة جدا . *hðh Alγrfħ Sγyrħ jdA .*
Beirut	هالأوضة كتير زغيرة . *hAlÂwDħ ktyr zγyrħ.*
Cairo	الأوضة دي صغيرة أوي . *AlÂwDħ dy Sγyrħ Âwy .*
Doha	هالغرفة واجد صغيرة . *hAlγrfħ wAjd Sγyrħ .*
Rabat	هاد الغرفة صغيرة بزاف . *hAd Alγrfħ Sγyrħ bzAf .*
Tunis	لبيت هذي صغيرة برشا. *lbyt hðy Sγyrħ bršA.*
Aleppo	هالغرفة كتير صغيرة . *hAlγrfħ ktyr Sγyrħ .*
Alexandria	الأوضة دى صغيرة جدا . *AlÂwDħ dý Sγyrħ jdA .*
Algiers	هاذ الغرفة صغيرة بزاف . *hAð Alγrfħ Sγyrħ bzAf .*
Amman	هاي الغرفة كتير صغيرة . *hAy Alγrfħ ktyr Sγyrħ .*
Aswan	الأوضة دي صغيرة خالص . *AlÂwDħ dy Sγyrħ xAlS .*
Baghdad	هاي الغرفة كولش صغيره . *hAy Alγrfħ kwlš Sγyrh .*
Basra	هاي الغرفة كلش صغيرة . *hAy Alγrfħ klš Sγyrħ.*
Benghazi	الدار صغيرة بكل . *AldAr Sγyrħ bkl.*
Damascus	هالغرفة صغيرة كتير . *hAlγrfħ Sγyrħ ktyr .*
Fes	هاد الغرفة صغيرة بزاف . *hAd Alγrfħ Sγyrħ bzAf.*
Jeddah	الغرفة دي مرا صغيرة . *Alγrfħ dy mrA Sγyrħ .*
Jerusalem	هاي الغرفة كتير صغيرة . *hAy Alγrfħ ktyr Sγyrħ .*
Khartoum	الغرفة دي صغيرة شديد . *Alγrfħ dy Sγyrħ šdyd .*
Mosul	الغرفه كلش صغيغي . *Alγrfh klš Sγyγy .*
Muscat	هالحجره وايد صغيرة . *hAlHjrh wAyd Sγyrħ .*
Riyadh	الغرفة صغيرة جدا . *Alγrfħ Sγyrħ jdA .*
Salt	هاي الغرفة كثير صغيرة . *hAy Alγrfħ kθyr Sγyrħ .*
Sanaa	الغرفة صغيرة قوي . *Alγrfħ Sγyrħ qwy.*
Sfax	البيت هذه ياسر صغيرة . *Albyt hðh yAsr Sγyrħ.*
Tripoli	الدار هادي صغيره هلبه . *AldAr hAdy Sγyrh hlbh .*

Figure 1: A sample of a 28-way parallel sentence extracted from CORPUS-25 including 5 sentences from CORPUS-5. The MSA and dialectal sentences are given along with their transliterations in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

3.1. Translation Guidelines

We provided the translators with the source sentences (out of context) and asked them to produce a translation that precisely reflects the source sentence without any assumption. We provided the translators with the following set of detailed guidelines:

- The translators were asked to use Arabic script, avoid any code-switching and to be internally consistent in spelling words. We did not provide them with any orthographic guidelines.
- Punctuation marks (such as periods, commas and question marks) that appear in the source sentence should remain in the Arabic dialect translation.
- Numbers written in letters should be translated into letters, while numbers written in digits should be kept as digits. For example, the translation of "six" is ستة *stħ*, while the translation of "6" is ٦.
- The translation of idioms should not be literal but reflect the meaning of the idioms instead.
- In the case where the gender (masculine vs. feminine) is not obvious in a source sentence, the masculine form should be used. For example, the English word *student* should be translated into Egyptian as طالب *TAlb* (masculine form in Arabic) not طالبة *TAlbħ* (feminine word in Arabic), unless the sentence contains a feminine form.[7] When the number (singular vs. plural) is not obvious in a sentence, the singular form should be used. For example, the English word *you* should be translated into Egyptian as انت *Ânt* not انتوا *ÂntwA*, as long as the sentence does not specify the plural form.
- Foreign words borrowed from English or French should be transliterated. For example, the French word *ordinateur (computer)* is commonly used in Tunisian Arabic, so it might be transliterated as اورديناتور *Awrdynatwr*. If the word has an equivalent in MSA, that is widely used in a certain dialect; this word should be translated into its MSA alternative. For example, the English word *program* should be translated as برنامج *brnAmj* not transliterated to بروجرام *brwjrAm*.

3.2. Corpus Analysis

The example in Figure 1 highlights the many lexical and morphological differences among the dialects of different cities. For example, the MSA word الغرفة *Alγrfħ* 'the room' was translated into البيت *Albyt* in Sfax, الأوضة *AlÂwDħ* in the Cairo, Alexandria and Aswan dialects and الدار *AldAr* in Tripoli and Benghazi dialects. While it was translated into the MSA-like form in other city dialects. This example shows the difference between various dialects, commonly treated as one big class of dialects such as Algerian, Moroccan (الغرفة *Alγrfħ)*, Tunisian (البيت *Albyt*) and Lybian (الدار *AldAr*).

In order to get an overall estimation of the similarity between the dialects of the cities covered in CORPUS-25 (in addition to MSA), we compute the Overlap Coefficient, representing the percentage of lexical overlap between the vocabularies for each dialect pair. The average pairwise similarity between the dialects in our dataset is 25.8% with

[7] We follow the choice made in producing the BTEC MSA version.

a standard deviation of 8.5% when MSA is included. When MSA is not included, the average similarity between the dialects is 26.3%. The most similar pair of dialects are those spoken in the cities of Amman and Jerusalem with an overlap of 54.4%. The least similar dialects are those of Sfax and Alexandria with a difference of 87.4%. The closest city dialect to MSA is Muscat dialect with an overlap score of 37.5%, and the most dissimilar one is the dialect of Sfax (lexical difference of 88.12%).

Overall, the lexical overlap between the dialects in our dataset is lower than the one reported in Bouamor et al. (2014). In the latter, the authors report high similarity scores between each dialect and Egyptian Arabic. This is explained by the fact that the translations were initially obtained from Egyptian which biased the lexical choices of the translators. This result justifies our decision to not use MSA as a starting point when building the MADAR corpus.

4. The MADAR Lexicon

In this section we present the structure of the MADAR lexicon and we describe the automatic and manual steps we followed in creating it.

4.1. Lexicon Structure

The MADAR lexicon is organized around *concept keys* which are defined in terms of triplets of words from English (En), French (Fr) and MSA. The multilingual triplets are intended to reduce ambiguity that comes from different senses of a particular word. For example, the English noun 'table' has a *furniture* sense and a *set of data* sense. But these two senses correspond to different MSA words طاولة *TAwlħ* and جدول *jdwl*, respectively. The latter of the MSA terms has other senses also, such as 'brook'. We plan to use these multilingual triplets to link to established large resources such as Wordnet (Fellbaum, 1998; Bentivogli et al., 2002) or Babelnet (Navigli and Ponzetto, 2012).

Each concept has a number of words associated with it. Each word is defined in terms of three aspects: its CODA orthography, its CAMEL Arabic Phonetic Inventory (CAPHI) phonology (Habash et al., 2018) and the various cities in which it is used. DA orthographic variations make it difficult for computational models to properly identify and reason about the words of a given dialect (Habash et al., 2012a). Hence, a conventional form for the orthographic notations is important to reduce sparsity and ambiguity. CODA is a set of guidelines and exception lists for Egyptian Arabic. Several efforts have extended them to cover other dialects (Jarrar et al., 2014; Zribi et al., 2014; Saadane and Habash, 2015; Turki et al., 2016; Khalifa et al., 2016). However, they focused on specific dialects and often made ad hoc decisions. In a recent effort, Habash et al. (2018) introduced a more unified set of guidelines and resources for DA orthography. They presented a common set of guidelines with enough specificity to help in creating dialect specific conventions as needed and applied them to the dialects of 25 Arab cities. In this work, we use these guidelines to build a CODAfied version of the MADAR Lexicon. [8]

Inspired by the International Phonetic Alphabet (IPA) and Arpabet (Shoup, 1980), CAPHI provides a system for transcribing all dialects of Arabic in a simple and user-friendly fashion, while still maintaining enough complexity to describe all meaningful phonological variation.[9] Figure 2 presents the full entry in the MADAR lexicon of the concept (very, très, جدّا *jdA*). The lexicon includes a small number of multi-word expressions, such as the Arabic multi-word expression representing the 'passport' concept in (passport, passeport, جواز سفر *jawAz safar*).

CODA	CAPHI	City Dialect
برشا *iršA*	b a r sh a	Tunis, Sfax
بزاف *bzAf*	b e z z aa f	Rabat, Fez, Algiers
بكل *bkl*	b i k k i l	Benghazi
جدا *jdA*	g i d d a n	Cairo, Alexandria
جدا *jdA*	j i d d a n	Jeddah, Khartoum, Riyadh
خالص *xAlS*	kh aa l i s.	Cairo, Alexandria, Aswan
شديد *šdyd*	sh a d ii d	Khartoum
قوي *qwy*	2 a w i	Cairo, Alexandria
قوي *qwy*	g a w i	Aswan, Sana'a
كثير *kθyr*	k i t ii r	Alexandria, Cairo
كثير *kθyr*	k t ii r	Beirut, Jerusalem, Damascus, Aleppo, Amman, Fez, Rabat
كثير *kθyr*	k th ii r	Amman, Salt
كثير *kθyr*	k th ii gh	Mosul
كثير *kθyr*	k a t ii r	Jeddah, Aswan, Khartoum
كثير *kθyr*	k i th ii r	Riyadh, Muscat
كلش *klš*	k u l l i sh	Basra, Baghadad
كلش *klš*	k e l l i sh	Mosul, Doha
مرة *mrħ*	m a r r a	Jeddah
هلبة *hlbħ*	h a l b a	Tripoli
عوم *ςwm*	3 oo m	Muscat
هوايا *hwAyA*	h w aa y a	Basra, Baghadad
واجد *wAjd*	w aa y i d	Basra, Baghadad, Doha
واجد *wAjd*	w aa j i d	Benghazi, Tripoli, Doha
واجد *wAjd*	w aa g i d	Muscat

Figure 2: MADAR Lexicon entries for concept (very, très, جدّا *jdA*).

Besides, each *concept key* is represented in a lemma and phrasal form. The lemma form is supplemented with its part-of-speech tag (POS). For Arabic, the *POS* is provided for the segmented form of the word on a clitical level. The phrasal form is a frequently used inflected form of the concept. For example, the concept of 'thanks' has a lemma form of *(thanks_NOUN merci_NOUN, šukr_NOUN* شكر*)*, while the phrasal form represents the Arabic word in its frequently used form of *šukrAã* شكرا. Also, the Arabic lemma form of the 'zoo' concept is *Hadiyqaħ_NOUN Al+_DET HayawAn_NOUN* حديقة ال + حيوان, while its phrasal

[8]A detailed description of CODA guidelines are available at: `http:\resources.camel-lab.com`.

[9]The complete CAPHI inventory is available at: `http:\resources.camel-lab.com`.

form is *Hadiyqaħ_NOUN Al+_DET HayawAnAt_NOUN* حديقة ال + حيوانات.

The MADAR lexicon contains a total of 1,045 concepts, which cover 88.0%, 86.4% and 85.5% of the lemma tokens in the English, French and MSA BTEC corpora respectively. Almost three-quarters of the concepts are for open classes.

4.2. Lexicon Concept Identification

Concept key identification relies on an automatic process that extracts *(English, French, Arabic)* related tuples from the BTEC parallel corpus. Tuples are then clustered based on their semantic similarity, such that each cluster represents a concept. The automatic process is followed by manual validation and fixing of errors resulting from the automatic process.

4.2.1. Automatic Extraction of Concept Keys

Data Preprocessing Since the concept triplet words are represented in terms of lemmas, we pre-process the parallel data to map it into the lemma space. For English, we use the Stanford POS tagger (Toutanova et al., 2003) and for French, we use Treetagger (Schmid, 1994). For Arabic, we use MADAMIRA (Pasha et al., 2014) to tokenize words into the D3 scheme, which separates all clitics from the basewords. Arabic tokenization is required as the clitics attached to basewords in Arabic, are typically represented as separate words in English and French. The most common examples are the proclitic definite article +ال *Al+* 'the', and the enclitic possessive pronouns, such as ه+ *+h* 'his'. The goal here is to harmonize the forms of the three languages to encourage better word alignment and concept extraction.

Triplet Extraction Our trilingual concept extraction approach focuses on collecting frequently used triplets. We align French-English, English-MSA, MSA-French pairs with GIZA++(Och, 2002) using the intersection symmetrization heuristic. Each word in an English sentence is aligned to words in the corresponding French and MSA sentences.

We address the triplet extraction problem as a task of collecting connected components from an undirected graph. Given three parallel English, French and MSA sentences, we represent words as nodes and alignments as edges in the graph. Nodes in an extracted component have to belong to the three languages strictly. Connected components are collected from all sentence pairs in the parallel aligned sentences, and each unique triplet is provided with a count representing the number of times the triplet is extracted from all the different parallel sentences. The output of the extraction method is a set of triplets sorted by their count. In Figure 3, we show an example of an aligned English, French and MSA parallel sentence.

Among the eight extracted connected components, four components constitute the triplets spanning the three languages: *(ce, the,* ال *Al+) , (acteur, actor,* ممثّل *mumaθil), (vraiment, really,* فعل *fiςl)* and *(merveilleux, marvelous,* رائع *rAŷiς).*

Concept Extraction Since several extracted triplets share some semantic similarity, we need to group the triplets into clusters such that each cluster represents a shared concept among the triplets. For example, the triplets *(bag, sac,* حقيبة *Haqiybaħ)*, *(bag, sac,* كيس *kiys)* and *(bag, baggage,* كيس *kiys)* represent the concept of a "bag" in the three languages. The concept can be represented by the triplet with the highest frequency. Our approach models this problem as a breadth-first traversal of an undirected graph where each triplet represents a node in the graph. An edge connects two triplets if they share two words from any of the three languages. For instance, we draw an edge between *(bag, sac,* حقيبة *Haqiybaħ)* and *(bag, sac,* كيس *kiys)* in Figure 4 since they share the English "bag" and the French "sac" constituents of the triplet.

We sort all triplets based on their frequency and apply a breadth-first traversal with a maximum depth of two, starting with the most frequent triplet. We iteratively repeat the breadth-first traversal starting with the next most frequent unvisited node, until all nodes are visited. The visited nodes in each traversal will constitute the cluster of a concept represented by the most frequent triplet.

Traversal with a depth of two (with respect to the starting node) was chosen empirically, as deeper levels showed some divergence from the main concept encompassed by triplets in the first two levels. In the undirected graph of Figure 4, we start with the highest frequency triplet (bag, sac, حقيبة *Haqiybaħ*) with a count of 134, and reach all neighboring triplets until a depth of two (shown in the left dotted square). The next most frequent triplet is (baggage, bagage, متاع *matAς*) with frequency of 102. We end up with two clusters representing the concepts of (bag, sac, حقيبة *Haqiybaħ*) and (baggage, bagage, متاع *matAς*).

4.2.2. Manual Validation of Concept Keys

The initial manual effort in building the lexicon involved carefully checking all the extracted concepts, correcting some cases and adding some missing entries. We identified four types of errors in the automatic lexicon construction approach we described above. First are preprocessing errors, mostly in the form of incorrect lemmatization. For example, يوقع *ywqς* 'to sign' was incorrectly lemmatized as وقع *waqaς* 'to fall' instead of وقّع *waq∼aς*. Second are missing alignment errors resulting from inherent linguistic differences. One example is the pronoun *I/je*, which is sometimes conjugated in Arabic as a verbal suffix. Since we use lemmas, the conjugated verbs are turned into their lemma form and that information is lost.

Third are multi-word expression (MWE) alignment errors. Since our approach did not address MWEs specifically, we had many cases of incomplete concept keys. For example, the English term 'really' in Figure 3 is incorrectly aligned to the Arabic term فعل *fiςl* 'act', while the correct align-

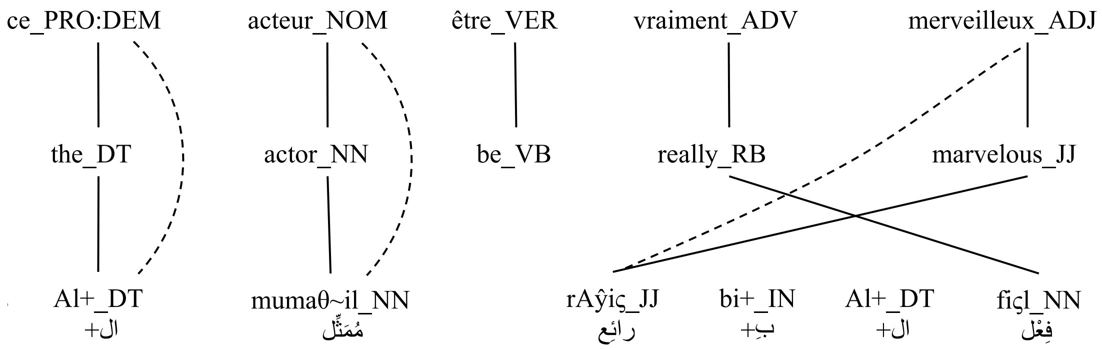

Figure 3: Alignment between French, English and MSA parallel sentences respectively. The non lemmatized forms of the three parallel sentences are: English: *The actor is really marvelous*; French: *Cet acteur est vraiment merveilleux*; MSA: المثل رائع بالفعل *Almumaθ∼il rAŷiς bAlfiςl*

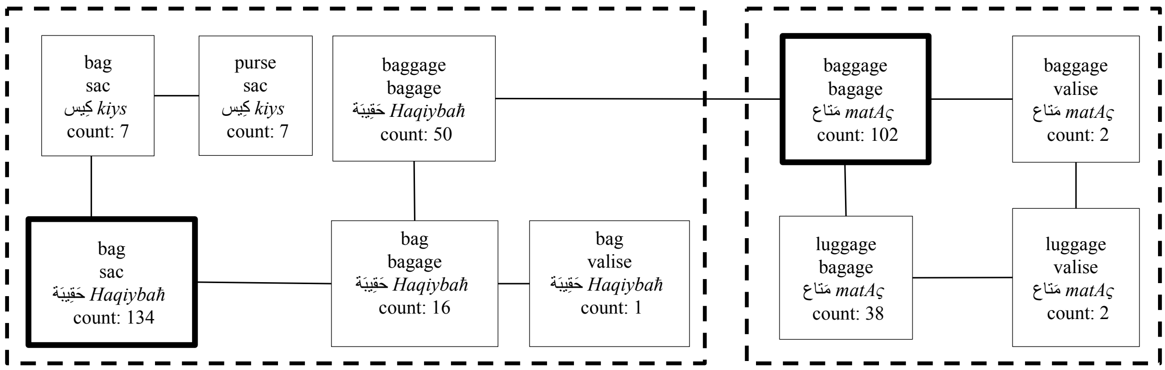

Figure 4: Concept extraction from aligned triplets. Each square represents a triplet with its English, French and Arabic terms and its count. Extracted concepts are indicated by the dotted square.

ment should be the MWE ب+ ال+ فعل *b+ Al+ fiEl*. Fourth are errors from very ambiguous words such as the word قد *qd* which means both *may be* and *certainly* depending on the aspect of the verb that follows it. This particular case ended up not included as the head of any single cluster despite being very frequent. Since we could not manually fix all the clusters, we targeted the top 1,000 or so clusters ranked by cluster-head frequency and recovered additional high-frequency words that were not properly identified automatically.

4.3. Lexicon Population

The lexicon population with dialectal entries proceeded in two steps. First, we automatically inserted entries extracted from a number of existing dictionaries; and then we manually validated, and extended them.

4.3.1. Automatic Lexicon Population

We transcribed a number of dialectal dictionaries: (i) The Karmous dictionary for Tunisian Arabic (Abdelatif, 2010), including around 3,800 words and several expressions and proverbs in Tunisian; (ii) the Moroccan Arabic Dialect textbook (Morocco, 2011), written by a team of language instructors who shared their collective experience gained by training thousands of Americans who lived and worked in Morocco. We also use the Tharwa lexicon (Diab et al., 2014), a four-way large-scale lexicon for dialectal Arabic, covering Egyptian and Levantine in addition to MSA and English; and the Iraqi dictionary from the LDC (Graff and Maamouri, 2009).

We attempted to populate our lexicon with as many entries by pivoting on English or French. These entries were not always in CODA-compliant form or had phonological representations that we could easily convert to CAPHI. We tried our best in this step to create CODA and CAPHI forms that are easy to edit and extend in the manual annotation step.

4.3.2. Manual Lexicon Population

The automatic lexicon population is followed by a large annotation effort, which involved 13 linguists who are from different regions of the Arab World. The lexicon is presented in a Google Sheet where every concept and its associated dialectal word forms are listed as shown in Figure 5.

There are two sections for every concept: The first section (marked in yellow cells) specifies the concept definition. The second section (marked in green cells) specifies the various dialect words. The concept definition consists of six columns including the concept ID (Concept_ID), its category, and in addition to the French, English and MSA

Concept_ID	Category	English	French	Standard Arabic	En-POS	Fr-POS	Ar-POS
139	Concept	car	voiture	سَيَّارة	NOUN	NOUN	NOUN
#=============							
Concept_ID	**Category**	**Dialect**	**Arabic CODA**	**CAPHI**	**Comments/Questions**		
139	AUTO	EGY	عربية	3 a r a b i y y a			
139	AUTO	LEV, IRQ, YEM	سيارة	s a y y aa r a			
139	AUTO	TUN	كرهبة	k r h b a			
139	ADD						

Figure 5: An example of an automatically populated concept, as presented to the linguists.

lemmas triplets, their corresponding POS tags (Fr-POS, En-POS, and Ar-POS). The dialectal word list consists of five columns including an identifier of the content of the row, and a category. The category could be: (a) **AUTO** for a word proposed by the automatic lexicon population described in 4.3.1.. It is not validated by a human, or (b) **ADD**: is an open slot provided to allow editors to add entries without inserting a new row. These values must be changed to **VALID**.

The column **Dialect** specifies the dialect of the entry. One or more region or city codes are provided per entry. The region code is provided instead of the city one for entries for which we do not have a city dictionary. For instance, the entry سيارة *syArħ* in Figure 5 was extracted from dictionaries covering these regional dialects Levantine (LEV), Iraqi (IRQ) and Yemeni (YEM). The linguists were asked to update this column with the corresponding specific city code. The code of each city is given in Table 1.

The linguists were provided with detailed guidelines on the steps to follow when editing and populating the lexicon. Each linguist was asked to:

- Read the concept definition carefully, clarify in his/her mind the exact meaning (this includes being aware of the full meaning and sub-meanings), and use the different translations and POS to help with this task.
- Scan the various AUTO entries provided for all regions. This might help him remember words that are possible candidates to add for the cities he/she is responsible for.
- Delete all entries that are NOT relevant to the cities he/she is responsible for.
- Apply the necessary changes for some entries that may need some minor fixes.
- Add new words that are not on the AUTO list.
- Think of more than one translation into his/her dialect and carefully specify the city.
- Use external informants to get more information for cities in his/her area if it is not his original city.
- Enter the CODA and CAPHI versions of each entry, using the guidelines provided.
- Make sure the Arabic CODA and CAPHI are correct for all the entries for their cities.
- Add the code names of the cities he/she is responsible for.
- Change the category to VALID once a row is fully validated.

Weekly meetings by the project PIs and a consulting lead linguist reviewed the progress of the linguists. At the time of writing this paper, the MADAR lexicon contained 47,466 dialectal words (average 1,899 per dialect). The average number of words per concept per dialect is 1.8. We are continuously working on quality checking, expanding and improving the coverage of the lexicon. [10]

5. Related Work

In the context of work on NLP, MSA has received the bulk of attention. There are lots of parallel and monolingual data collections, richly annotated collections (e.g., treebanks), sophisticated tools for morphological analysis and disambiguation, syntactic parsing, etc. (Habash, 2010). Even for languages other than Arabic, the integration of dialectal variation in NLP applications is rather rare. One interesting exception is the work of Scherrer (2012) on Swiss German dialects.

Very recently, automatic DA processing has attracted a considerable amount of research in NLP (Shoufan and Alameri, 2015), facilitated by the newly developed monolingual and multilingual dialectal corpora and lexicons. Several mono-dialectal corpora covering different Arabic dialects were built and made available. Al-Badrashiny et al. (2014) compiled a large dialect-identified corpus of DA from several Egyptian sources, but with a large presence of MSA. In a related effort, McNeil and Faiza (2011) built a four-million-word corpus of Tunisian Spoken Arabic. Various other research work resulted in multidialectal non parallel corpora at different scales (Zaidan and Callison-Burch, 2011; Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Salama et al., 2014; Jeblee et al., 2014; Al-Shargi et al., 2016; Zaghouani and Charfi, 2018).

[10] The latest version of the lexicon is available for browsing online at http://nlp.qatar.cmu.edu/madar/.

As for dialect-to-dialect parallel corpora, Bouamor et al. (2014) presented the first small-scale 7-way parallel corpus covering several dialects in addition to MSA, and English, all translated from Egyptian sentences. The fact that Egyptian was chosen as a starting point affected the quality of the translation. The sentences produced were biased by the use of some Egyptian expressions that might be accepted in other dialects, but a native would not produce naturally. The same concern applies to the 6-way parallel PADIC corpus used in Meftouh et al. (2015), as all translations were derived from DA or MSA. When developing CORPUS-5 and CORPUS-25, we avoided such priming effects by asking translators to produce translations starting from English or French based on their preferences. However, most of these efforts focus primarily on a number of varieties corresponding generally to those spoken in major cities (Cairo, Amman, Baghdad, etc.), or study different dialects independently.

Unlike MSA, DA has a small number of printed bilingual or monolingual dictionaries. Thus, building a DA lexicon with varying degrees of coverage and linguistic complexity has been the aim of several research efforts. The LDC built the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al., 2002) and the Iraqi Arabic Morphological Lexicon (IAML) (Graff and Maamouri, 2009), two mono-dialectal lexica, that were used in developing the Egyptian and Iraqi versions of the CALIMA morphological analyzer (Habash et al., 2012b).

A notable multi-dialectal lexicon is the one built in the *Arabic Variant Identification Aid (AVIA)* project. This lexicon covers the seven Arabic dialects spoken in the following cities: Al-Ain (United Arab Emirates), Baghdad (Iraq), Jeddah (Saudi Arabia), Jerusalem (Palestinian Arabic), Kuwait City, Doha (Qatar) and Sana'a (Yemen). Another significant effort is Tharwa (Diab et al., 2014), a 4-way English, MSA, Egyptian, Levantine lexicon with rich linguistic annotation. Our lexicon is similar to Tharwa in that we also use CODA compliant lemma forms. However, the MADAR lexicon includes phonetic modeling via the CAPHI representation. Also, our lexicon covers more regional and city dialects (25 city dialects) compared to Tharwa (two dialects only). *The Dialects of Arabic* project at the University of Manchester recently made publicly available a database of Arabic dialects that include a mix of words and sentences in their phonological forms covering samples from 15 countries in the Arab World (Matras and others, 2017).

To the best of our knowledge, our work is the first effort aiming at building large scale and fine-grained dialectal Arabic resources (corpora and lexicon) mapped to their English, French and MSA versions.

6. Conclusion and Future Work

We presented two resources: the MADAR Corpus and MADAR Lexicon. The first is a large scale parallel corpus created by translating selected sentences in the travel domain into 25 Arabic city dialects. The second is a lexicon of 1,045 entries covering the same 25 Arabic cities. These resources are the first of their kind in terms of the breadth of coverage and fine granularity.

In the future, we plan to extend both resources in terms of number of cities. We also plan to expand the lexicon with more entries. The MADAR Corpus and Lexicon will be used to create three enabling technologies and applications that are necessary to support future research in Arabic NLP: dialect identification, machine translation and morphological analysis.

Acknowledgements

We would like to thank our dedicated linguists who contributed in building the MADAR lexicon: Linda Alamir, Feryal Albrehi, Shumool Albuainain, Gazella Ben Sreiti, Jamila El-Gizuli, Dihia Gareche, Fatma Ghailan, Anissa Jrad, Reham Marzouk, Mohammad Abuodah, Salim Al-Mandhari and Aous Mansouri. We also would like to thank *Ramitechs* for the translation services, and the UStar Consortium for providing us with the English version of BTEC.

This publication was made possible by grant NPRP 7-290-1-047 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

7. Bibliographical References

Abdelatif, K. (2010). *Dictionnaire le Karmous du Tunisien*.

Abu-Melhim, A.-R. (1991). Code-switching and Linguistic Accommodation in Arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized Dialectal Arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 30–38, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for moroccan and sanaani yemeni arabic. In *10th Language Resources and Evaluation Conference (LREC 2016)*.

Bassiouney, R. (2009). *Arabic Sociolinguistics*. Edinburgh University Press.

Benmamoun, E. (2012). Agreement and Cliticization in Arabic Varieties from Diachronic and Synchronic Perspectives. In Reem Bassiouney, editor, *Al'Arabiyya: Journal of American Association of Teachers of Arabic*, volume 44-45, pages 137–150. Georgetown University Press.

Bentivogli, L., Pianta, E., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A Multidialectal Parallel Corpus of Arabic. In *Proceedings* *of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland.

Brustad, K. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Cotterell, R. and Callison-Burch, C. (2014). A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245.

Darwish, K. (2014). Arabizi Detection and Conversion to Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224, Doha, Qatar.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3782–3789, Reykjavik, Iceland.

Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Fellbaum, C. (1998). Wordnet: An electronic lexical database. *The MIT Press*.

Graff, D. and Maamouri, M. (2009). Iraqi Arabic Morphological Lexicon (IAML) Version 6.5. Linguistic Data Consortium.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic transliteration. In Abdelhadi Soudi, et al., editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, chapter 2, pages 15–22. Springer.

Habash, N., Diab, M., and Rambow, O. (2012a). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718, Istanbul, Turkey.

Habash, N., Eskander, R., and Hawwari, A. (2012b). A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada.

Habash, N., Khalifa, S., Eryani, F., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Haeri, N. (1991). Sociolinguistic Variation in Cairene Arabic: Palatalization and the qaf in the Speech of Men and Women.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a corpus for Palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27.

Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., and Oflazer, K. (2014). Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). Egyptian Colloquial Arabic Lexicon.

Matras, Y. et al. (2017). Database of Arabic Dialects. The University of Manchester.

McNeil, K. and Faiza, M. (2011). Tunisian Arabic corpus: Creating a written corpus of an unwritten language. In *Workshop on Arabic Corpus Linguistics (WACL)*.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *The Proceedings of the 29th Pacific Asia conference on Language, Information and Computation*.

Morocco, P. C. (2011). *Moroccan Arabic*.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Och, F. (2002). *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Germany.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Saadane, H. and Habash, N. (2015). A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, page 69, Beijing, China.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating Dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria.

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014). YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland.

Scherrer, Y. (2012). *Generating Swiss German Sentences from Standard German: a Multi-dialectal Approach*. Ph.D. thesis, University of Geneva, Switzerland.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Shoufan, A. and Alameri, S. (2015). Natural Language Processing for Dialectical Arabic: A Survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China.

Shoup, J. E. (1980). Phonological Aspects of Speech Recognition. *Trends in speech recognition*, pages 125–138.

Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180, Edmonton, Canada.

Turki, H., Adel, E., Daouda, T., and Regragui, N. (2016). A Conventional Orthography for Maghrebi Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia.

Watson, J. C. (2007). *The Phonology and Morphology of Arabic*. Oxford University Press.

Zaghouani, W. and Charfi, A. (2018). ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of ACL 2011*, Portland, Oregon, USA.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and 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*, pages 49–59. Association for Computational Linguistics.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze Khmekhem, M., Hadrich Belguith, L., and Habash, N. (2014). A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland.