Morphologically Annotated Corpora for Seven Arabic Dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan

Faisal Alshargi,[⋆] Shahd Dibas,[‡] Sakhar Alkhereyf,[†] Reem Faraj,[†]
Basmah Abdulkareem,[†] Sane Yagi,[‡] Ouafaa Kacha,[‡] Nizar Habash,[*] Owen Rambow[§]
[⋆]Universität Leipzig, Germany [‡]University of Jordan, Jordan [†]Columbia University, USA
[*]New York University Abu Dhabi, UAE [§]Elemental Cognition, USA
alshargi@informatik.uni-leipzig.de, shahddibas@hotmail.com, sakhar@cs.columbia.edu,
nizar.habash@nyu.edu, owen.rambow@gmail.com

Abstract

We present a collection of morphologically annotated corpora for seven Arabic dialects: Taizi Yemeni, Sanaani Yemeni, Najdi, Jordanian, Syrian, Iraqi and Moroccan Arabic. The corpora collectively cover over 200,000 words, and are all manually annotated in a common set of standards for orthography, diacritized lemmas, tokenization, morphological units and English glosses. These corpora will be publicly available to serve as benchmarks for training and evaluating systems for Arabic dialect morphological analysis and disambiguation.

1 Introduction

As Arabic dialects (DA) become more widely written in social media, there is increased interest in the Arabic NLP community to have annotated corpora that will allow us to both study the dialects linguistically, and to create systems that can automatically process dialectal text. There have been important efforts to create relatively large corpora for Egyptian (Maamouri et al., 2014), Palestinian (Jarrar et al., 2014), and Emirati Arabic (Khalifa et al., 2018). While these resources are very helpful for single dialects, the problem is that there are many dialects, and in fact it is often unclear what to count as separate dialects (for example, the subdialects of Levantine). Therefore, we present a different approach in this paper: we annotate seven dialects, but with relatively smaller corpora (most around 30,000 words). Some of the dialects are closely related (Jordanian and Syrian), others are more distant (Moroccan). We use the same annotation methodology for all dialects: same guidelines, same processing steps, and same annotation file format. This makes our effort an ideal starting point for experimenting with using multidialectal resources to create and train NLP tools. The dialects we consider are Taizi Yemeni (YE.TZ)[1], Sanaani Yemeni (YE.SN), Saudi Najdi (SA.NJ), Jordanian (JOR), Syrian Damascene (SY.DM), Iraqi Baghdadi (IR.BG), and Moroccan Rabati (MA.RB) Arabic.

The paper is structured as follows. We start with a review of relevant literature (Section 2). We then summarize some linguistic facts about DA in general (Section 3) and subsequently present each of our seven dialects in Section 4, summarizing the corpora used and some interesting facts specific to each dialect. Section 5 then presents our annotation methodology. We then briefly discuss morphological analyzers, and conclude.

2 Related Work

Data Collections There have been several data collections centered on Arabic dialects, specifically spoken Arabic. A very useful resource is the Semitisches Tonarchiv at the University of Heidelberg in Germany.[2] We have included two Yemeni transcriptions from this resource in our YE.TZ and YE.SN corpora. Khalifa et al. (2016) is a large collection of over 100M words of a number of Arabic dialect, although the majority is from the Gulf. Bouamor et al. (2018) created a large corpus with parallel data text from 25 Arab cities. Further data collections include (Al-Amri, 2000) which has not yet been digitized for use in NLP research.

Annotated Corpora There are few annotated corpora for dialectal Arabic: the Levantine Arabic Treebank (specifically Jordanian) (Maamouri et al., 2006), the Egyptian Arabic Treebank (Maamouri et al., 2014), Curras, the Pales-

[1]The abbreviations we use intend to capture the country name and the city or region name when applicable.

[2]http://www.semarch.uni-hd.de

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 137–147
Florence, Italy, August 1, 2019.

tinian Arabic annotated corpus (Jarrar et al., 2014), the Gulf Arabic Annotated corpus (Khalifa et al., 2018), Syrian, Jordanian dialectal corpora (Bouamor et al., 2014; Harrat et al., 2014), a small effort on Sanaani and Moroccan (AlShargi et al., 2016) (which this paper builds on), and SUAR (Al-Twairesh et al., 2018), a morphologically annotated corpus for Najdi and Hijazi which is semi-automatically annotated using the MADAMIRA tool (Pasha et al., 2014) and subsequently manually checked. Additionally, Voss et al. (2014) present a corpus of Moroccan dialect which has been annotated for language variety (code switching). Several of these efforts have followed the approach of Curras (Jarrar et al., 2014), which consists of around 70,000 words of a balanced genre corpus. The corpus was manually annotated using the DIWAN tool (Alshargi and Rambow, 2015), which we also use. The annotation in Curras is done by first using a morphological tagger for another Arabic dialect, namely MADAMIRA Egyptian (Pasha et al., 2014), to produce a base that was then corrected or accepted by a trained annotator.

Other NLP Resources for Dialectal Arabic The effort to annotate corpora in context is a central step in developing morphological analyzers and taggers (Eskander et al., 2013; Habash et al., 2013). However, other notable approaches and efforts that do not use annotated corpora have focused on developing specific resources manually or semi-automatically, e.g., the Egyptian Arabic morphological analyzer (Habash et al., 2012b) which is built upon the Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002), the multi-dialectal dictionary Tharwa (Diab et al., 2014), or extending MSA analyzers and resources (Salloum and Habash, 2014; Harrat et al., 2014; Boujelbane et al., 2013).

Linguistic Studies There are many theoretical and descriptive linguistic studies for the dialects we work on: Yemeni dialects (Watson, 1993, 2002), Najdi (Ingham, 1994), Gulf Arabic dialect (Holes, 1990), Jordanian (Bani-Yasin and Owens, 1987), Moroccan (Harrell, 1962), Syrian (Cowell, 1964), and Iraqi (Erwin, 1963); not to mentions comparative studies across dialects and MSA (Holes, 2004; Brustad, 2000). We make extensive use of such studies as part of the design of our annotation guidelines.

3 Dialects: Linguistic Facts

In this section we present some general facts and phenomena shared across different dialects. In subsequent subsections, we present our dialects in more detail and commenting on the corpus sources.

Dialects and MSA Arabic dialects share many commonalities with Classical Arabic and Modern Standard Arabic (MSA). All variants of Arabic are morphologically complex as they include rich inflectional and derivational morphology that is expressed in two ways: namely, via templates and affixes. Furthermore, they contain several classes of attachable clitics. However, the dialects as a class differ in consistent ways from MSA, and they differ amongst each other. In fact, the differences between MSA and Dialectal Arabic (DA) have often been compared to those between Latin and the Romance languages (Chiang et al., 2006). The principal morpho-syntactic difference between DA and MSA is the loss of productive case marking, and nunation (*tanween*) on nouns, and mood on imperfective verbs.

Dialectal Variations Differences among the dialects are found on all levels of linguistic description, i.e., phonology, morphology, syntax, and the lexicon. We summarize three phonological and three morphological salient examples in Table 1 for our dialects: the pronunciation of MSA /q/ written ق *q*,[3] MSA /ʤ/ written ج *j* and MSA /k/ written ك *k*; and the various forms of the future, progressive and possessive particles.

From a lexical point of view, there are many words that have different meanings across dialects. For example, the word ماشي *mA$y* /maːʃi/ is 'no' in YE.SN and MA.RB, 'yes/ok' in SY.DM and JOR, and 'walking' in SA.NJ. Another example is the word صافي *SAfy* /sˁaːfi/ which means 'enough' in MA.RB, but 'pure' in the other dialects and MSA. Some cases show subtle differences in meaning, e.g., خدام *xdAm* /xaddaːm/ means 'employee' generically in MA.RB, but it has a more specific and negative connotation in YE.TZ and YE.SN, namely 'enslaved servant'. While the above cases are all homonyms (homophones and homographs), there are instances of

[3] We represent the Arabic words in Arabic script and in the Buckwalter transliteration (in italics) (Habash et al., 2007). When needed, we present the IPA (in /.../). The English gloss is added in single quotes.

Phenomenon	MSA	YE.TZ	YE.SN	SA.NJ	JOR	SY.DM	IR.BG	MA.RB
Pronunciation of ق *q*	/q/	/q/	/g/	/g/ or /dz/	/g/ or /ʔ/	/ʔ/	/g/	/q/ or /g/
Pronunciation of ج *j*	/ʤ/	/g/	/ʤ/	/ʤ/	/ʒ/	/ʒ/	/ʤ/	/ʤ/
Pronunciation of ك *k*	/k/	/k/	/k/	/k/ or /ts/	/k/ or /ʧ/	/k/	/k/ or /ʧ/	/k/
Future Particle	+س *s+* سوف *swf*	+ش *$+* اش *A$*	+ع *E+* عد *Ed* +ش *$+* +ي *y+*	+ب *b+*	+ح *H+* رح *rH*	+ح *H+* رح *rH*	+ح *H+* رح *rH* راح *rAH*	+غ *g+* غادي *gAdy*
Progressive Particle	ϕ	+ب *b+*	+ب *b+*	قاعد *qAEd* جالس *jAls*	+ب *b+*	+ب *b+* عم *Em*	+د *d+* قاعد *qAEd*	+ك *k+* +ت *t+*
Possessive Particle	ϕ	تبع *tbE* حق *Hq*	تبع *tbE* حق *Hq*	حق *Hq*	تبع *tbE*	تبع *tbE* تاع *tAE*	مال *mAl*	+د *d+* ديال *dyAl*

Table 1: Cross-dialectal and MSA variants in some phonological and morphological phenomena

homophones that have different meanings in different dialects. For example the utterance /faʒr/ can mean 'morning' in YE.TZ (written as فجر *fjr*), or 'poverty in YE.SN (written as فقر *fqr*). The YE.SN pronunciation of فجر *fjr* is /faʒr/; and the YE.TZ pronunciation of فقر *fqr* is /faqr/.

There are also cases of the same meaning being expressed in different ways, e.g., 'spoon' is ملعقة *mlEqp* in MSA, metathesized معلقة *mElqp* in JOR and SY.DM, and خاشوقة *xA$wqp* in IR.BG.

Dialectal Orthography Since Arabic dialects do not have spelling standards, several previous efforts on Arabic dialect annotations (Maamouri et al., 2014; Jarrar et al., 2014; Khalifa et al., 2018) contributed to a movement that lead to the creation of a common Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2012a; Zribi et al., 2014; Habash et al., 2018). We also follow this approach to map from any *spontaneous* orthography in our data to CODA. The spirit of CODA is to define a common and consistent approach to spelling DA words that acknowledges their etymological and historical relationship with MSA and CA, but also maintains their uniqueness and independence. For example, if a DA word has an MSA cognate containing ق *q*, then its CODA spelling will use ق *q* even if the dialectal pronunciation is different. In contrast, DA morphemes are spelled in a way to reflect their DA uniqueness. For example the SY.DM word حنفيق *Hnfyq* /ħanfiːʔ/ 'we will wake up' is a cognate of MSA سنفيق *snfyq* /sanfiːqu/: the future marker reflects the dialectal morphology and is not spelled as in MSA, but the stem is spelled as in MSA and thus the ق *q* does not reflect the dialectal pronunciation.

4 Dialect-Specific Corpora

Until recently, Arabic was mostly written in Modern Standard Arabic (MSA) and Classical Arabic, while written DA was rare. One early source of written dialectal Arabic are textbooks for learning an Arabic Dialect intended for non-Arabic speakers. Furthermore, sometimes spoken language has been recorded and transcribed. However, owing to the advent of the internet and its rapid growth among Arabic speaking populations, written materials in DA are now more accessible and easy to obtain than they were in the past. These written materials are typically informal written conversations among participant or traditional folk literature like short stories, poems, prose, thoughts and song. These texts can be found in online forums, blogs, and postings on social media networks. All of the our dialectal corpora consist of sources of various genres, collected from both online and print materials in order to cover many of the aspects of these dialects. Each of the YE.TZ, SA.NJ, IR.BG, JOR corpora has 30K words, while the YE.SN has 32K words, SY.DM has 35k words and MA.RB has 20k words. It should be noted that the data collected from the internet was written in Arabic characters, using "spontaneous" orthography since there are no orthographic standards for DA. The Roman alphabet sentence were transcribed from the textbooks into the Arabic alphabet using CODA. All examples presented in the rest of this section are in CODA except where specified otherwise.

4.1 Taizi Corpus (YE.TZ)

Sources The YE.TZ written data was collected manually from different resources such as forums,

blogs, and social media networks. With reference to spoken data, half of the oral interviews were recorded and transcribed manually by the annotators, the remaining oral interview transcripts are taken from the Semitisches Tonarchiv (Section 2). The data includes wise anecdotes, proverbs, stories, poems, songs and dialogues.

Phonology and Orthography A distinguishing feature of YE.TZ is that MSA ج *j* /ʤ/ is pronounced as /g/, e.g., جمل *jml* 'camel' /gamal/, and that MSA ق *q* /q/ retains its pronunciation. In that regard, CODA spellings were straightforward.

Morphology Similar to a number of other dialects but unlike MSA, negation is expressed as an enclitic ش *$* 'not', e.g., يدخلش *ydxl+$* 'he does not enter'. The vocative particle is expressed as the proclitics يا *yA* 'Oh' and وا *wA* 'Oh', or as an the enclitic اه *Ah* as in أمّاه *AmAh* 'my mother'. The verbal proclitic قا *qA* 'already', which corresponds to MSA قد *qd*, frequently appears with past verbs, e.g., قا عملنا *qA EmlnA* 'we have already done that'.

Lexicon There are many open-class words that make YE.TZ different from MSA and other dialects, e.g., زقوة *zqwp* 'shrewd', زكن *zkn* 'order', and قراع *qrAE* 'breakfast'. Some words have MSA meanings that differ from YE.TZ, e.g., شل *$l* 'take' and بزّ *bz* 'take'. YE.TZ has a number of loanwords from English that underwent Arabization, e.g., سجارة *sjArp* 'cigarette', and كتلي *ktly* 'kettle'.

4.2 Sanaani Corpus (YE.SN)

Sources The social texts were taken from a Sanaani Radio Station program called مسعد ومسعدة *msEd wmsEdp*, which addressed social issues and problems of the community. The oral interview transcripts were taken from the Semitisches Tonarchiv (Section 2). The interviews describe daily life, history and lifestyle in Sanaa. Folktales describing traditional stories handed down in Sanaa are taken from internet forums. Collections of wisdom sayings and tales of the famous wise man of Yemen "Ali walad Zaid" are taken from internet websites. Other texts were taken from social media, and include political events in Yemen, Sanaani jokes, religious sermons and transcripts that discuss the Sanaani dialect in MSA.

Phonology and Orthography MSA ق *q* /q/ is pronounced /g/ in YE.SN, including in religious contexts. For example, the word قمر *qmr* 'moon' is pronounced /gamar/. This variation is not unique to YE.SN and other dialects such as IR.BG and JOR have it as well. This /g/ is often spontaneously spelled as ق *q*, which is consistent with CODA guidelines. A particularly marking phenomenon in YE.SN is the devoicing and emphasis of some instances of word-medial /d/, e.g., غدوة *gdwp* 'tomorrow' is pronounced /ɣutˁwa/ and as a result may be written spontaneously as غطوة *gTwp*.

Morphology As shown in Table 1, there are four future particles in YE.SN: ع+ *E+*, عد *Ed*, ش+ *$+*, ي+ *y+*. While ع+ *E+* may be used with 1st, 2nd, or 3rd person conjugated verb, the rest are only used with 1st person singular conjugated verbs.

Lexicon YE.SN has some distinguishing closed class words, such as prepositions قفى *qfY* 'behind' and شق *$q* 'next', and numbers like ستات *stAt* 'six', and هطعش *hTE$* 'eleven'. There are also some Turkish loanwords, e.g., ساني *sAny* 'direct' and كريك *kryk* 'shovel'.

4.3 Najdi Corpus (SA.NJ)

Sources The SA.NJ corpus was collected from different sources that represent different genres: forums, poetry, jokes and tweets. We collected different posts from the Saudi web forum eqla3.com, including personal narratives (mainly sarcastic) and discussions. We also collected Najdi poems from the late twentieth century, mainly written by the contemporary Najdi poets Khalid AlFaisal, Mohammed bin Ahmed AlSudairy and Saad Bin Jadlan. We manually collected Najdi jokes from various online resources. And finally, on Twitter, we searched for distinctive Najdi keywords such as حنا *HnA* 'we', قروشة *qrw$p* 'inconvenience', and منيب *mnyb* 'I'm not'.

Phonology and Orthography As Table 1 shows, there are a number of phonological alternations in SA.NJ. The /dz/ variant of ق *q* /q/ and /ts/ variants of ك *k* /k/ are rather restricted in their usage. And unlike MSA, SA.NJ shows no distinction between the pronunciation of MSA etymological ض /dˁ/ and ظ /ðˁ/. These phenomena affect spontaneous orthography and had to be addressed in the CODA annotations.

Morphology One marking morphological feature of SA.NJ (and other Gulf Arabic dialects) is

the use of negation circumfix ب+ .. ـش ما+ *mA+ .. +b*, as in مانيب *mAnyb* 'I am not' (spontaneously, often written as منيب *mnyb*). Similar constructions exist in other dialects but are more productive, e.g. Egyptian ب+ .. ـش ما+ *mA+ .. +$* negates verbs in addition to pronouns. Unlike most DA and like MSA, SA.NJ retains some tanween (nunation). For example: أنا قايلٍ لك *>nA qAylK lk* /ʔana ga:ylin lak/ 'I said (active participle) to you'. However, as in MSA, the nunation is rarely written. Some morphological phenomena are becoming very rare, e.g., the use of تس *ts* for 2nd person singular feminine pronominal enclitic is dying out among younger people and merging with the masculine form ك *k*.

Lexicon SA.NJ has some distinguishing words such as أبخص *>bxS* 'more expert', كفو *kfw* 'good', and دافور *dAfwr* 'nerd' There are many borrowed words from English compared to borrowings from Turkish or Persian. For instance, the verb يفلم *yflm* is borrowed form English 'film' and means 'to act dramatically'.

4.4 Jordanian Corpus (JOR)

Sources The corpus includes written as well as spoken data. The written materials were drawn from internet sources, such as, forums, blogs, and social media. They include informal conversations among participant or traditional folk literature like short stories, poems, prose, memoirs, and songs. As for spoken data, oral interviews and observations were recorded and transcribed by the annotators. Nearly 20 informants were interviewed by the researchers. Older as well as uneducated people are included in order to ensure the authenticity of the data. The JOR data included a mix of subdialects that reflect the multiplicity of DA forms, including markedly Palestinian as well as Jordanian variants. For this reason, we refer to this corpus simply as JOR.

Phonology and Orthography In some JOR sub-dialects, as with IR.BG, MSA ك *k* is affricated to /ʧ/, e.g., كلب *klb* /ʧalb/ 'dog'. ق *q* also realizes in two forms as /g/ and /ʔ/. Some of these phenomena results in different spontaneous spellings that are then normalized during annotation.

Morphology JOR's 2nd person feminine singular pronominal clitic has two alternations depending on the sub-dialect: كي *ky* /ki/ and ك *k* /ik/. Examples include شفتكي *$ftky* or شفتك *$ftk* 'I saw you'; however when following a vowel, both become كي *ky* /ki/, e.g. شافوكي *$Afwky* 'they saw you'. Negation is marked with the enclitic ش *$*; such as, باسويش *bAswy$* 'I do not do'.

Lexicon Some JOR words are from Syriac, e.g., شوب *$wb* 'hot', and بكير *bkyr* 'early in the morning'. Other words are borrowed from Turkish, e.g., دغري *dgry* 'straightforward' and درابزين *drAbzyn* 'ladder'. Some words that were borrowed from English underwent some morpho-phonological changes. For example, كوريدور *kwrydwr* 'corridor', فرمت *frmt* 'format', and بلك *blk* 'to block somebody'.

4.5 Syrian Corpus (SY.DM)

Sources The written data was collected manually from different online written resources such as forums, blogs, and social media networks. Among the data, there were anecdotes, proverbs, stories, some poems, songs and dialogues.

Phonology and Orthography SY.DM has a glottal stop phoneme /ʔ/ that is a cognate with either MSA Hamza (ء & } > < ' ئ إ أ ؤ) or MSA Qaf ق *q*. In most spontaneous SY.DM orthography, the two forms are distinguished in a manner similar to CODA guidelines. A few exceptions include the word هلأ *hl>* 'now' which in CODA is written as هلق *hlq* highlighting its etymological link to هالوقت *hAlwqt* 'this time'. Less common spelling variations include the devoicing of ج *j* /ʒ/ to /ʃ/, which may be reflected in spontaneous orthography, e.g., نجتمع *njtmE* /niʒtmiʕ/ 'we meet' may appear as نشتمع *n$tmE* /niʃtmiʕ/.

Morphology A distinction of SY.DM (and North Levantine) compared to South Levantine and a number of other dialects is the absence of the negation enclitic ش *$*. SY.DM makes use of a number of future particles in free distribution (See Table 1). The progressive particle عم *Em* can only be used to indicate active progression at the moment, while the progressive proclitic ب+ *b+* has a wider range from habitual to progressive.

Lexicon As with JOR, some SY.DM words were originally Syriac, e.g., شوب *$wb* 'hot', or براني *brAny* 'outer'. Other words are borrowed from Turkish, e.g., دغري *dgry* 'straightforward'. Some words encountered major semantic shifts, e.g., طز *Tz* comes from Turkish *tuz* 'salt', then shifting to mean 'something unimportant', and eventually

'good riddance'. Other words were found to be borrowed from French, e.g., ديكور *dykwr* 'decor' and جاتو *gAtw* 'gateaux', and from Persian like سرسري *srsry* 'bad man'. Markedly SY.DM expressions include حربوق *Hrbwq* /ħarbuːʔ/ 'shrewd'.

4.6 Iraqi corpus (IR.BG)

Sources The materials of the IR.BG corpus were obtained from social media websites, blogs and other online sources. The sources contain posts on political, social, and religious issues that touch upon the daily life of the Iraqi people. The sources include blogs, e.g., different sarcastic posts with a witty sense of humor gathered from the Iraqi blog شلش العراقي *l AlErAqy*, and short essays with commentary and views that sharply criticize loss in traditional values and morals in the Iraqi society after 2003. Proverbs, common sayings, and famous expressions were also collected from online blogs and forums.

Phonology and Orthography Some instances of MSA ك *k* appear as /tʃ/ in IR.BG, e.g., كانت *kAnt* 'she was' /tʃaːnat/. Some of these cases appear in spontaneous orthography as تش *t$* or even چ *J*/ج *j* (mostly due to Persian spelling influences). Some instances of MSA /q/ are pronounced as /g/, e.g., فوق *fwq* 'above' /foːg/. Some of these cases appear in spontaneous orthography as گ *G* or ك *k*, also due to Persian influences.

Morphology A strong marker of IR.BG is the progressive proctlitc د+ *d+*, e.g., شدتسوق؟ *$dtswq?* 'what are you driving?'. IR.BG also has three future particles: راح *rAH*, رح *rH*, and ح+ *H+*, which seem to be in free variation.

Lexicon The IR.BG lexicon has some distinguishing words such as أطوخ >*Twx* 'little darker', and آني |*ny* 'I'. IR.BG has many loanwords from Kurdish, Persian, and Russian, e.g., Kurdish كاكه *kAkh* 'mister', Persian قنداغ *qndAg* 'very weak tea or hot water and sugar', and Russian إستكان <*stkAn* 'a spindle-shaped tea cup'.

4.7 Moroccan Corpus (MA.RB)

Sources The corpus includes comments from the Moroccan news website hespress.com that have to do with sports, cinema, and education policy. The materials from forums include advice on social, religious, and economic issues. The oral interviews are transcriptions of people telling stories, most of which are events from their lives. The folktales come from a Moroccan website that reprinted stories originally published in an encyclopedia of traditional Moroccan folktales. The textbook examples include many basic greetings and expressions, as well as sample dialogues. The blog posts range in topic, but include relationship advice, recipes, and philosophical musings. The humor includes both short and long jokes from a few Facebook pages and one other website.

Phonology and Orthography Most MA.RB consonants are pronounced like their MSA equivalents; however, there are exceptions: dental consonants in MSA have become alveolar, so MSA ث *v* /θ/, ذ * /ð/, and ظ *Z* /ðˤ/, are pronounced /t/, /d/, and /dˤ/, respectively in MA.RB. Such issues naturally interact with spontaneous orthography and are annotated as per CODA guidelines.

Morphology Among the set of dialects discussed here, MA.RB has the most distinct set of morphological features, such as its future, progressive and possessive particles (see Table 1). Like other North African dialects, and unlike MSA, MA.RB uses the prefix ن+ *n+* for imperfect first person singular, and distinguishes first person plural by adding the plural suffix +وا *+wA*. Interestingly the imperfect first person singular in MA.RB looks like the imperfect first person plural in MSA and numerous other dialects. Finally, the perfect second person singular masculine and feminine both use the suffix تي *ty*, which corresponds to the feminine suffix in other DA.

Lexicon MA.RB has a number of loanwords from Berber, French and Spanish; and many speakers code-switch between Moroccan and French or Spanish. Examples include French فورماج *fwrmAj* 'cheese', and بورتابل *bwrtAbl* 'mobile phone'; and Spanish سمانة *smAnp* 'week', and بابور *bAbwr* 'ship'.

5 Annotation Process

Process Overview To create new morphological annotated corpora, we follow (AlShargi et al., 2016)'s basic approach: we utilize the DIWAN tool (Alshargi and Rambow, 2015) to build and annotate the seven DA corpora discussed above. The project team consists of:

1. a project manager,
2. dialect leads for each dialect, and

	Gloss	*to him*	*and I will not go / and not going*			*this letter*		*I will write*	
MSA	**Ortho**	إليه	أذهب	ولن		الرسالة	هذه	سأكتب	
	Lemma	<ilaY	*ahab	lan		risAlap	h‘*A	katab	
	Morph	<lY +h	A+ *hb +ϕ	w+ ln		Al+ rsAl +p	h*h	s+>+ ktb +ϕ	
	Prefix	-	IV1S	CONJ		DET	-	FUT_PART+IV1S	
	Stem	PREP	IV	NEG_PART		NOUN	DEM_PRON_FS	IV	
	Suffix	PRON_3MS	IVSUFF_MOOD:S	-		NSUFF_FEM_SG +CASE_DEF_ACC	-	IVSUFF_MOOD:I	
YE.TZ	**Raw**	لو	شسرحش	وما		الجواب	أذه	شكتب	
	CODA	له	شاسرحش	وما		الجواب	اذه	شاكتب	
	Lemma	li	saraH	mA		jawAb	Aa*ah	katab	
	Morph	l +h	$+A+ srH +$\phi$+$	w+ mA		Al+ jwAb	A*h	$+A+ ktb +$\phi$	
	Prefix	-	FUT_PART+IV1S	CONJ		DET	-	FUT_PART+IV1S	
	Stem	PREP	IV	NEG_PART		NOUN	DEM_PRON_MS	IV	
	Suffix	PRON_3MS	IVSUFF_SUBJ:1S+NEG_PART	-		-	-	IVSUFF_SUBJ:1S	
YE.SN	**Raw**	له	شسيرش	وما		الرسالة	تيه	عدكتب	
	CODA	له	شاسيرش	وما		الرسالة	تيه	عد اكتب	
	Lemma	li	sAr	mA		risAlap	tayh	katab	
	Morph	l +h	$+A+ syr +$\phi$+$	w+ mA		Al+ rsAl +p	tyh	Ed#+A+ ktb +ϕ	
	Prefix	-	FUT_PART+IV1S	CONJ		DET	-	FUT_PART#+IV1S	
	Stem	PREP	IV	NEG_PART		NOUN	DEM_PRON_FS	IV	
	Suffix	PRON_3MS	IVSUFF_SUBJ:1S+NEG_PART	-		NSUFF_FEM_SG	-	IVSUFF_SUBJ:1S	
SA.NJ	**Raw**	له	رايح	ومنيب		هالرساله		بكتب	
	CODA	له	رايح	ومانيب		هالرسالة		باكتب	
	Lemma	li	rAyH	AnA		risAlap		katab	
	Morph	l +h	rAyH	w+m+ Any +b		h+Al+ rsAl +p		b+A+ ktb+ϕ	
	Prefix	-	-	CONJ+NEG_PART		DEM_PART+DET		FUT_PART+IV1S	
	Stem	PREP	ADJ	PRON_1S		NOUN		IV	
	Suffix	PRON_3MS	-	NEG_PART		NSUFF_FEM_SG		IVSUFF_SUBJ:1S	
JOR	**Raw**	ليه	رايح	ومنا		الرسالة	هاذي	كتب	رح
	CODA	ليه	رايح	ومنا		الرسالة	هاذي	اكتب	رح
	Lemma	li	rAH	mnA		risAlap	hA*iy	katab	raH
	Morph	l +h	rAyH	w+ mnA		Al+ rsAlp	hA*y	A+ ktb +ϕ	rH
	Prefix	-	-	CONJ		DET	-	IV1S	-
	Stem	PREP	ADJ	NEG_PART		NOUN	DEM_PRON_FS	IV	FUT_PART
	Suffix	PRON_3MS		-		NSUFF_FEM_SG	-	IVSUFF_SUBJ:1S	-
SY.DM	**Raw**	لعندو	روح	رح	وما	هالرسالة		أكتب	رح
	CODA	لعنده	اروح	رح	وما	هالرسالة		اكتب	رح
	Lemma	Eind	rAH	raH	mA	risAlap		katab	raH
	Morph	l+ End +h	A+ rwH +ϕ	rH	w+ mA	h+Al+ rsAl +p		A+ ktb +ϕ	rH
	Prefix	PREP	IV1S	-	CONJ	DEM_PART+DET		IV1S	-
	Stem	NOUN	IV	FUT_PART	NEG_PART	NOUN		IV	FUT_PART
	Suffix	POSS_PRON_3MS	IVSUFF_SUBJ:1S	-	-	NSUFF_FEM_SG		IVSUFF_SUBJ:1S	-
IR.BG	**Raw**	له	اروح		وما	الرسالة	هاي	كتب	رح
	CODA	له	اروح		وما	الرسالة	هاي	اكتب	رح
	Lemma	li	rAH	mA		risAlap	hAy	katab	raH
	Morph	l +h	A+ rwH +ϕ	w+ mA		Al+ rsAlp	hAy	A+ ktb +ϕ	rH
	Prefix	-	IV1S	CONJ		DET	-	IV1S	-
	Stem	PREP	IV	NEG_PART		NOUN	DEM_PRON_FS	IV	FUT_PART
	Suffix	PRON_3MS	IVSUFF_SUBJ:1S	-		NSUFF_FEM_SG	-	IVSUFF_SUBJ:1S	-
MA.RB	**Raw**	ليه	تمشي	غاديش	وما	الرساله	هاد	نكتب	غادي
	CODA	ليه	تمشي	غاديش	وما	الرسالة	هاد	نكتب	غادي
	Lemma	li	m$aY	gAdy	mA	risAlap	hAd	ktab	gAdy
	Morph	l +h	n+ m$y +$\phi$	gAdy +$	w+ mA	Al+ rsAlp	hAd	n+ ktb +ϕ	gAdy
	Prefix	-	IV1S	-	CONJ	DET	-	IV1S	-
	Stem	PREP	IV	FUT_PART	NEG_PART	NOUN	DEM_PRON_FS	IV	FUT_PART
	Suffix	PRON_3MS	IVSUFF_SUBJ:1S	NEG_PART	-	NSUFF_FEM_SG	-	IVSUFF_SUBJ:1S	-

Table 2: An annotation example from DIWAN for Modern Standard Arabic, Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan Arabic dialects. All the sentences have the same meaning: ‘I will write this letter and not go to him’. The table is presented in a right-to-left direction. **Raw** represents a spontaneous word spelling. **CODA** represents the conventional orthography we use. **Lemma** shows the diacritized lemma form; this is the only line where we show diacritics. **Morph** represent the sequence of prefixes, the stem, and the sequence of suffixes. **Prefix**, **Stem**, and **Suffix** show the part of speech tags for the components of the word shown in the **Morph** line.

Error Type	Dialects	Word	gloss	Error	Correction
Null Subject	SA.NJ	آمر |*mr*	order	+|mr/CV+	+|mr/CV+(null)/CVSUFF_SUBJ:2MS
	YE.TZ	أصابحك >*SAbHk*	fight	>/IV1S+SAbH/IV+k/IVSUFF_DO:2MS	>/IV1S+SAbH/IV +(null)/IVSUFF_SUBJ:1S +k/IVSUFF_DO:2MS
Ta-Marbuta	SY.DM	جعبتي *jEbty*	pouch	+jEb/NOUN+p/NSUFF_FEM_SG +y/POSS_PRON_1S	+jEb/NOUN+t/NSUFF_FEM_SG +y/POSS_PRON_1S
Case	SY.DM	بالسقف *bAlsqf*	roof	b/PREP+Al/DET+sqf/NOUN +(null)/CASE_DEF_GEN	b/PREP+Al/DET+sqf/NOUN+

Table 3: Examples of annotation errors found during error analysis: null morphemes should be added; ta-marbuta is a common source of errors; case should never be annotated for the dialects

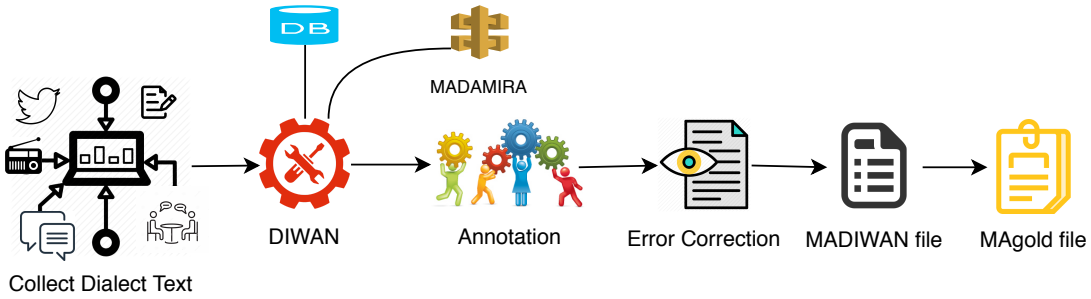

Figure 1: Steps to creating a new annotated corpus for a dialect

3. annotators.

The dialect leads verify the annotators' work, and the project manager organizes and monitors the flow of the progress of everyone using the tool in the project.

Annotation Steps First, the dialect leads collect the corpus text from different resources like social media, forms, websites, etc. The next step is to develop dialect-specific annotation guidelines, including the CODA specification for normalized orthography. The dialect leads then train the annotators before annotation starts. The leads follow the annotator's work. The annotations are not approved until the dialect leads check them. Wrong annotations are sent back to the annotator for correction. After the first round of annotation is done, we perform a second round of error checking, using both manual inspection and scripts that check for coherent annotations. The result is a DIWAN file which includes the correct annotation for the entire corpus. In the last step, we automatically reformat the annotations into a format which is best suited for computational purposes; we perform a third round of error checking for format errors, which we fix automatically. Figure 1 shows these steps.

Morphological Features Annotated The DIWAN interface assists human annotators in annotating each token with morphological and semantic information, including the following fields:

- The CODA spelling of the raw token.
- The lemma, or the citation form, of the token.
- The morphemes of the word (prefixes, stem, suffixes) and their part-of-speech (POS). The stem is marked by the symbol # on either side.
- The English gloss of the word.
- Features indicating proclitics and enclitics.
- Features indicating word POS, functional number and gender (Alkuhlani and Habash, 2011), and aspect.

The annotation for one sentence in different dialects is shown in Table 2. This is not actually a sentence from our corpora, of course; we have chosen it to illustrate the annotation.

Error Correction Linguistic annotation is carried out manually. In order to guarantee high levels of accuracy and precision, we performed extensive error checking and correction. After annotating the seven different corpora, the annotated words were compiled in the form of linguistic codes in either one file or separate files to be

checked and corrected by a second reviewer. This form of error checking cannot of course identify annotation errors in context (for example, a noun is misidentified as a verb); instead, this approach is efficient at finding impossible annotations. Examining the data demonstrated that the most challenging part for the annotators was the suffixes part, especially when there are long and complicated words. Some examples indicating the errors are listed below in Table 3.

Distribution of Resources All created resources will be freely available for research purposes from Columbia (`http://innovation.columbia.edu`).

6 Conclusion and Future Work

We presented a collection of morphologically annotated corpora for seven Arabic dialects, collectively covering over 200,000 words. All corpora were manually annotated in a common set of standards for orthography, diacritized lemmas, tokenization, morphological units and English glosses. These corpora will be publicly available to serve as benchmarks for training and evaluating systems for Arabic dialect morphological analysis and disambiguation.

In future work, we will use these resources to train morphological taggers as described in (Eskander et al., 2016). We also plan to extend the collection of dialect to include additional less studied varieties following the lead of efforts such as Bouamor et al. (2018). We also plan to expand towards different historical and literature based varieties of Arabic.

7 Acknowledgments

This work is supported by the Air Force Research Laboratory (AFRL) under a grant administered by Ball Aerospace. Alkhereyf is supported by the KACST Graduate Studies program. The views expressed here are those of the authors and do not reflect the official policy or position of the U.S. Department of Defense or the U.S. Government We also would like to thank all the anonymous reviewers for their insightful and valuable comments and suggestions.

References

Abd Al-Salam Al-Amri, editor. 2000. *Texts in Sanani Arabic. O*. Harrassowitz, Wiesbaden, Germany.

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, et al. 2018. Suar: Towards building a corpus for the Saudi dialect. *Procedia computer science*, 142:72–82.

Sarah Alkuhlani and Nizar Habash. 2011. A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Portland, Oregon, USA.

Faisal AlShargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow:. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Faisal Alshargi and Owen Rambow. 2015. Diwan:a dialectal word annotation tool for Arabic. In *In: Proceedings of WANLP 2015 - ACL-IJCNLP, 2015*.

Raslan Bani-Yasin and Jonathan Owens. 1987. The phonology of a northern jordanian arabic dialect. *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, 137(2):297–331.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Rahma Boujelbane, Mariem Ellouze Khemekhem, and Lamia Hadrich Belguith. 2013. Mapping rules for building a Tunisian dialect lexicon and generating corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 419–428.

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

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Mark Cowell. 1964. *A Reference Grammar of Syrian Arabic*. Georgetown University Press, Washington, D.C.

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 2014. Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3782–3789, Reykjavik, Iceland.

Wallace Erwin. 1963. *A Short Reference Grammar of Iraqi Arabic*. Georgetown University Press, Washington, D.C.

Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic extraction of morphological lexicons from morphologically annotated corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1032–1043, Seattle, Washington, USA. Association for Computational Linguistics.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016. Creating resources for dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3455–3465, Osaka, Japan. The COLING 2016 Organizing Committee.

Nizar Habash, Mona T. Diab, and Owen Rambow. 2012a. Conventional orthography for dialectal Arabic. In *LREC*.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 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. Association for Computational Linguistics.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaili. 2014. Building resources for Algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*.

Richard Harrell. 1962. *A Short Reference Grammar of Moroccan Arabic: With Audio CD*. Georgetown classics in Arabic language and linguistics. Georgetown University Press.

Clive Holes. 1990. *Gulf Arabic*. Croom Helm Descriptive Grammars. Routledge, London / New York.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Bruce Ingham. 1994. *Najdi Arabic*. John Benjamins.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 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, Doha, Qatar. Association for Computational Linguistics.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. *CoRR*, abs/1609.02960.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Hanaa Kilany, Hassan Gadalla, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, and Cynthia McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *LREC*, pages 2348–2354.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *LREC*, volume 14, pages 1094–1101.

Wael Salloum and Nizar Habash. 2014. Adam: Analyzer for dialectal Arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, 26(4):372–378.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014. Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2249–2253, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1086.

Janet Watson, editor. 1993. *A syntax of Sanani Arabic*. O.Harrassowitz, Wiesbaden, Germany.

Janet Watson. 2002. *The Phonology and Morphology of Arabic*. Oxford University Press.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *LREC*, pages 2355–2361.