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 multdialectal resources to create and train NLP tools. The dialects we consider are Taizi Yemeni (YE.TZ), 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 (specifcally 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
tinian Arabic annotated corpus (Jarrar et al., 2014), the Gulf Arabic Annotated corpus (Khaliﬁa 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 mention 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 q, MSA /h/ written j and MSA /k/ written 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 مامى (صايف /mqaʃ/) 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 صايف /mqaʃ/ which means ‘enough’ in MA.RB, but ‘pure’ in the other dialects and MSA. Some cases show subtle diﬀerences in meaning, e.g., مامى xdAm /xadda:m/ 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

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.
homophones that have different meanings in different dialects. For example the utterance /faqr/ can mean ‘morning’ in YE.TZ (written as JOR and SY.DM, and ‘poverty in YE.SN (written as the MSA, but the stem is spelled as in MSA and thus the dialectal morphology and is not spelled as in MSA, but the stem is spelled as in MSA and thus the 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 /l/by/ is pronounced as /g/ by. e.g., ممل 'camel' /gamal/, and that MSA ق q /l/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 مهمل $ ‘not’, e.g., سيلخت ‘he does not enter’. The vocative particle is expressed as the proclitics ا يا ‘O h’ and ا يا ‘Oh’, or as an enclitic ا يا ‘my mother’. The verbal proclitic ا ‘already’, which corresponds to MSA ج qd, frequently appears with past verbs, e.g., ا ا ‘we have already done that’.

**Lexicon**

There are many open-class words that make YE.TZ different from MSA and other dialects, e.g., رفعة ‘shovel’, رك ‘knife’, ج ‘order’, and قَام ‘breakfast’. Some words have MSA meanings that differ from YE.TZ, e.g., سأ ‘I am’, and بز ‘take’. YE.TZ has a number of loanwords from English that underwent Arabization, e.g., التل ‘cigarette’, and كلمة ‘kettle’.

**4.2 Sanaani Corpus (YE.SN)**

**Sources**

The social texts were taken from a Sanaani Radio Station program called مسعود ومسمى 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 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 ج q /l/q/ is pronounced /g/ in YE.SN, including in religious contexts. For example, the word فر qmr ‘moon’ is pronounced /qmar/. 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 /l/, e.g., في ‘tomorrow’ is pronounced /yut/ and as a result may be written spontaneously as /Tw/.

**Morphology**

As shown in Table 1, there are four future particles in YE.SN: + E, + Ed, + $, + +. While + E can 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 for ‘behind’ and حيث ‘next’, and numbers like ستة stÅ ‘six’, and هفط tÅ ‘eleven’. There are also some Turkish loanwords, e.g., دAny ‘direct’ and كريك krik ‘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 31a8.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 Jadal. We manually collected Najdi jokes from various online resources. And finally, on Twitter, we searched for distinctive Najdi keywords such as حنا Hna ‘we’, كرز qrw$ ‘inconvenience’, and صبي 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 ج q /l/q/ and /ls/ variants of ع k /k/ are rather restricted in their usage. And unlike MSA, SA.NJ shows no distinction between the pronunciation of MSA etymological ج /d/ and ج /g/. 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...
Lexicon  SA.NJ has some distinguishing words such as : أخص kfw ‘good’, and دافور dAfwr ‘nerd’ There are many borrowed words from Arabic compared to borrowings from Turkish or Persian. For instance, the verb يفلم yfljm 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 sub-dialects 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 د k is affricated to /j/ / , e.g., كله klh /j/abl ‘dog’. Q also realizes in two forms as /j/ and /j/ / . 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 /k/ and د k /kl/. Examples include شفكتي $fisky or شفتكي $fik ‘I saw you’; however when following a vowel, both become شافكتي $Afisky ‘they saw you’. Negation is marked with the enclitic $; such as, لستي baSwy$ ‘I do not do’.

Lexicon Some JOR words are from Syriac, e.g., شوب Swb ‘hot’, and يبكر bkyr ‘early in the morning’. Other words are borrowed from Turkish, e.g., خ druAbzyn ‘ladder’. Some words that were borrowed from English underwent some morpho-phonological changes. For example, كوريدور kwrydwir ‘corridor’, وفمت frmt ‘format’, and بلک 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 /h/ that is a cognate with either MSA Hamza (و) or Qaf  ق. In most spontaneous SY.DM orthography, the two forms are distinguished in a manner similar to CODA guidelines. A few exceptions include the word ماء h2l ‘now’ which in CODA is written as هاوله hAq be highlighting its etymological link to هاوله hAq ‘this time’. Less common spelling variations include the devoicing of ج j /j/ to /j/, which may be reflected in spontaneous orthography, e.g., نتاج njtmiQ ‘we meet’ may appear as نتاج njtmiQ.

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 Ed  can only be used to indicate active progression at the moment, while the progressive proclitic + has a wider range from habitual to progressive.

Lexicon As with JOR, some SY.DM words were originally Syriac, e.g., شوب Swb ‘hot’, or يراني البراني brAny ‘outer’. Other words are borrowed from Turkish, e.g., صحاري dgry ‘straightforward’. Some words encountered major semantic shifts, e.g., تز تز comes from Turkish تز ‘salt’, then shifting to mean ‘something unimportant’, and eventually
‘good riddance’. Other words were found to be borrowed from French, e.g., dykw ‘decor’ and گآsw ‘gateaux’, and from Persian like ۱ر۱ ‘bad man’. Markedly SY.DM expressions include حروب ۱ر۱ ‘shrewd’.

4.6 Iraqi corpus (IR.BG)
Sources The materials of the IR.BG corpus were obtained from media websites, blogs, and other online sources. The sources contain posts on political, religious, and social 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 لاش الن، 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 ۲k appear as ۲ timings in IR.BG, e.g., ۲کAnt ‘she was’ / fantasies. Some of these cases appear in spontaneous orthography as ۲ش or even ۲ج (mostly due to Persian spelling influences). Some instances of MSA ۲g are pronounced as ۲yg, e.g., فvq ‘above’ /fog/. Some of these cases appear in spontaneous orthography as ۲G or ۲k, also due to Persian influences.

Morphology A strong marker of IR.BG is the progressive proclitic ۵+ in IR.BG, e.g., ۵کانت ‘what are you driving?’. IR.BG also has three future particles: ۵رA, ۵رH, and ۵+ ۵ح, which seem to be in free variation.

Lexicon The IR.BG lexicon has some distinguishing words such as ۵دژ ‘little darker’, and ۵ن ‘I’. IR.BG has many loanwords from Kurdish, Persian, and Russian, e.g., Kurdish کک ‘mister’, Persian ۵دنم ‘very weak tea or hot water and sugar’, and Russian ۵کآن ‘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 other websites.

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 ۵v /θ/, ۵* /ð/, and ۵Z /d/ are pronounced /t/, /l/, and /d/, 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 ۵+ ۵ن for imperfect first person singular, and distinguishes first person plural by adding the plural suffix ۵+ ۵نA. 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 ۵ن, 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 فورم ‘cheese’, and Arabic ۵بهر ‘mobile phone’; and Spanish ۵امن ‘week’, and Arabic ۵بهر ‘ship’.

5 Annotation Process
Process Overview To create new morphologically 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
<table>
<thead>
<tr>
<th>Gloss</th>
<th>to him</th>
<th>and I will not go / and not going</th>
<th>this letter</th>
<th>I will write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ortho</td>
<td>&lt;i&gt;aY</td>
<td>لاحب</td>
<td>lan</td>
<td>nS</td>
</tr>
<tr>
<td>Lemma</td>
<td>l+iY</td>
<td>ماسم</td>
<td>saraH</td>
<td>mA</td>
</tr>
<tr>
<td>Stem</td>
<td>PREP</td>
<td>CONJ</td>
<td>IV</td>
<td>IV</td>
</tr>
<tr>
<td>Suffix</td>
<td>PRON,MS</td>
<td>NSUFF,MODA:S</td>
<td>I VSUFF,MODA:1</td>
<td>I VSUFF,MODA:1</td>
</tr>
</tbody>
</table>

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** shows the diacritized lemma form; this is the only line where we show diacritics. **Morph** represents 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.

---

**DIWAN** shows the diacritized lemma form; this is the only line where we show diacritics. **Morph** represents 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.
<table>
<thead>
<tr>
<th>Error Type</th>
<th>Dialects</th>
<th>Word</th>
<th>gloss</th>
<th>Error</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Subject</td>
<td>SA.NJ</td>
<td>آنر</td>
<td>order</td>
<td>+[mr/CV+]</td>
<td>+[mr/CV+(null)]/CVSUFF.SUBJ:2MS</td>
</tr>
<tr>
<td>YE.TZ</td>
<td>أسامة</td>
<td></td>
<td>fight</td>
<td>&gt;/IV1S+SAbH/IV+ CV+</td>
<td>&gt;/IV1S+SAbH/IV+(null)/IVSUFF.SUBJ:1S+ k/IVSUFF.DO:2MS</td>
</tr>
<tr>
<td>Ta-Marbuta</td>
<td>SY.DM</td>
<td>إرب</td>
<td>pouch</td>
<td>+jEB/NOUN+p/NSUFF.FEM.SG</td>
<td>+y/POSS_PRON.1S</td>
</tr>
<tr>
<td>Case</td>
<td>SY.DM</td>
<td>بالسقف</td>
<td>roof</td>
<td>NPREP+Al/DET+sqf/NOUN</td>
<td>b/PREP+Al/DET+sqf/NOUN</td>
</tr>
</tbody>
</table>

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

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 DI-WAN 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 corpus, 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 annotating the suffixes, especially when there are long and complicated words. Some examples indicating the errors are listed in Table 3.

Distribution of Resources All created resources will be freely available for research purposes from a Columbia University website (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 was 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


