Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, Owen Rambow

Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic

Faisal Al-Shargi[*], **Aidan Kaplan**[†], **Ramy Eskander**[‡], **Nizar Habash**[§], **Owen Rambow**[‡]
[*]Universität Leipzig, Germany; [†]Yale University, USA; [‡]Columbia University, USA
[§] New York University Abu Dhabi, United Arab Emirates
alshargi@informatik.uni-leipzig.de, aidan.kaplan@yale.edu, rnd2110@columbia.edu
nizar.habash@nyu.edu, rambow@ccls.columbia.edu

Abstract

We present new language resources for Moroccan and Sanaani Yemeni Arabic. The resources include corpora for each dialect which have been morphologically annotated, and morphological analyzers for each dialect which are derived from these corpora. These are the first sets of resources for Moroccan and Yemeni Arabic. The resources will be made available to the public.

Keywords: Corpus, Arabic Dialects, Morphological Analysis

1. Introduction

Traditionally, Arabic dialects have been mainly spoken, and only rarely written. Most written Arabic has been Classical Arabic or Modern Standard Arabic (MSA), and therefore most natural language processing (NLP) tools address these two variants of Arabic. However, this situation is currently undergoing change. The expanding use of social media and electronic forms of written communication has been accompanied in the Arab world by an increase in the production of written dialect. As a result, there is a growing interest in NLP tools for written Arabic dialects, for example from companies that would like to perform sentiment data mining on regional web forums to determine what consumers think about their products. However, few such NLP tools for Arabic dialects exist. An important factor that contributes to the lack of NLP tools for the dialects is that there are many distinct dialects and few annotated corpora. This paper presents coordinated resource creation efforts for two Arabic dialects: Moroccan (MOR) and the Sanaani Dialect of Yemen (YEMS). For each dialect, we present a new morphologically annotated corpus and a morphological analyzer. These are the first such resources for Moroccan and for Yemeni, to our knowledge.

This paper is structured as follows. We start with related work (Section 2.). We then present some linguistic facts for our two dialects (Section 3.). Section 4. contains details about our two corpora, and we sketch our annotation scheme in Section 5.. We discuss the creation of the morphological analyzers in Section 6., and then conclude. The conclusion also contains information about how to obtain our resources.

2. Related Work

There has been a fairly large amount of descriptive work on Moroccan Arabic, with prominent publications including a reference grammar (Harrell, 1962) and a dictionary (Harrell et al., 2004), as well as in-depth work on Moroccan syntax (Brustad, 2000). For Yemeni Arabic, there has also been a fair amount of work in theoretical and descriptive linguistics (Jastrow, 1984; Behnstedt, 1985; Abdullah, 1991; Watson, 1993; Naïm-Sanbar, 1994; Al-Iryani, 1996; Behnstedt, 2006).

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[1] under the direction of Prof. Werner Arnold. We have included one Yemeni transcription from this resource in our Yemeni corpus (see Section 4.). Further data collections include (al Salam Al-Amri, 2000).

There are few *annotated* corpora for dialectal Arabic. We note three corpora in particular: the Levantine Arabic Treebank (specifically Jordanian) (Maamouri et al., 2006), the Egyptian Arabic Treebank (Maamouri et al., 2014) and Curras, the Palestinian Arabic annotated corpus (Jarrar et al., 2014). Additionally, (Tratz et al., 2014) present a corpus of Moroccan dialect which has been annotated for language variety.

Our work follows the work of Curras (Jarrar et al., 2014), which consists of around 43,000 words of a balanced genre corpus. The corpus was manually annotated using the DIWAN tool (Al-Shargi 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. Since Arabic dialects do not have spelling standards, the team working on Curras followed previous efforts to create conventional orthographies (or CODA) for the dialect they worked on (Habash et al., 2012a; Zribi et al., 2014). We also follow this approach and define CODAs for MOR and YEMS.

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). Other notable approaches and efforts 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), multi-dialectal corpora (Bouamor et al., 2014; Smaïli et al., 2014), the Gulf Arabic corpus (Khalifa et al., 2016) or extending MSA analyzers and resources (Salloum and Habash, 2014; Smaïli et al., 2014; Boujelbane et al., 2013).

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

3. Linguistic Facts

Dialectal Arabic poses many challenges for NLP. Arabic in general is a morphologically complex language which includes rich inflectional morphology, expressed both templatically and affixationally, and several classes of attachable clitics. For example, the Moroccan Arabic (MOR) word وغيكتبوها *w+γa+y-ktb-uw+hA*[2] 'and they will write it' has two proclitics (+و *w+* 'and' and +غ *ga+* 'will'), one prefix -ي *y-* '3rd person masculine imperfective', one suffix -و *-uw* 'plural' and one pronominal enclitic +ها *+hA* 'it/her'. In Sanaani Yemeni Arabic (YEMS), the equivalent word is وعيكتبوها *w+ςa+y-ktb-uw+hA* 'and they will write it'; it has two proclitics (+و *w+* 'and' and +ع *ς+* 'will'), one prefix -ي *y-* '3rd person imperfective', one suffix -و *-uw* 'masculine plural' and one pronominal enclitic +ها *+hA* 'it/her'. In both dialects, the word is considered an inflected form of the lemma *katab* 'write [lit. he wrote]'.

Both Moroccan (MOR) and Sanaani Yemeni Arabic (YEMS) share many similarities with other Arabic dialects. For example, they lack the inflectional categories for case and mood that are found in Modern Standard Arabic (MSA). Moroccan Arabic (MOR) is a part of the Maghrebi dialect group, and so it is especially similar to dialects such as Algerian and Tunisian. Sanaani Arabic (YEMS) is a variety of Yemeni Arabic, and it shares features with other varieties of the Arabian Peninsula.

In the following subsections, we discuss some of the distinctive features of MOR and YEMS.

3.1. Phonology

Many vowels which are short in other dialects are reduced to schwa or deleted completely in MOR. Vowels which are long in other dialects are often pronounced semi-long in MOR, and vowel length is typically not contrastive. Most MOR consonants are pronounced like their MSA equivalents; however, there are a few sound changes. Dental consonants in MSA have become alveolar consonants, so MSA /θ, ð, ðˤ/, represented in MSA by the letters ث *θ*, ذ *ð*, and ظ *Ď*, correspond to [t, d, dˤ] in MOR, as if these sounds were spelled ت *t*, د *d*, and ض *D* respectively. Words that have /q/ in MSA, spelled ق *q* may be pronounced in MOR with [q] or [g] (a sound that is absent in MSA), and some speakers, especially in Fes, use [ʔ], as if it were spelled أ '. Additionally, [g] appears in some words that have /ʒ/ in MSA, spelled ج *j*, such as [gləs] 'he sat' in MSA /ʒalasa/, and it appears in words of non-Arabic origin, such as [garo] 'cigarette'.

YEMS also has sound changes that make it differ from MSA. MSA /q/ has become [g] in YEMS as well, included in religious contexts. For example, the MSA word قمر *qmr* 'moon' is pronounced in YEMS with an initial [g]. Word medially, MSA /d/, spelled د *d*, is usually pronounced in YEMS as [tˤ], as if it were spelled ط *T*. For example, the YEMS word meaning 'tomorrow' (in MSA /ɣudwa/ غدوة *gdwħ*) becomes [ɣutˤwa], as if it were spelled غطوة *γTwħ*.

3.2. Morphology

Distinguishing features in MOR include some clitics, such as ك *ka-*, which indicates progressive or indicative present-tense verbs, and غ *ga-*, which indicates future tense. Like other North African dialects, and unlike MSA, MOR uses the prefix ن *n-* for present tense first person singular, and distinguishes first person plural by adding the plural suffix وا *-uwA*. Additionally, past tense second person singular masculine and feminine both use the suffix تي *-tiy*, which corresponds to the feminine suffix in other varieties of Arabic.

YEMS maintains the MSA gender distinction in the 3rd and 2nd person plural pronouns, but it makes no gender distinction in the 1st person singular. The presentative particle ذا *ðA* is linked with the preceding pronoun, giving هو ذا *hw ðA* 'he is', and هي ذا *hy ðA* 'she is'. There are three main future particles in YEMS: ش *š*, ع *ς*, ي *y*. The particles ش *š* and ي *y* are only used with 1st person, while the future particle ع *ς* may be used with 1st, 2nd, or 3rd person. In 1st person singular, a د *d* is added after the ع *ς* to form, for example, عد اكتب *ςd Aktb* 'I will write'. Both relative pronouns اللّي *Ally* and الّذي *Alðy* are used, depending on the region of Sana'a.

3.3. Syntax

MOR has several unique constructions. For example, there is a possessive particle ديال *dyAl* 'of', which often is used instead of the *idaafa* construction. ديال *dyAl* has a higher degree of grammaticalization than similar forms in other dialects, such as Levantine تبع *tabaς*. In contrast to تبع *tabaς*, ديال can reduce to د *d* and cliticize, e.g. بزاف دالفلوس *bzAf dAlfluws* 'a lot of money'. Another distinguishing construction is the use of واحد *wAHd* plus the determiner ال *Al-* as an indefinite article, e.g. واحد البلاصة *wAHd AlblASaħ* 'a place; this one place'.

YEMS has a few distinctive constructions. The common negative particles are مع *maς* and ماشي *mAšy*. The existential is expressed using the particle به *bih*. The conditional particle in YEMS is لاشي *lAšy* 'if'. In a question, this combines with the existential به *bih* to form شي به؟ *šy bih?* 'is there?'.

3.4. Lexicon

MOR has a number of loanwords from Berber, French and Spanish, and many speakers code-switch between Moroccan and French or Spanish.

YEMS has some unique words in closed classes, such as prepositions قفي *qfý* 'behind' and شق *šq* 'next', صلى *Slý* 'toward', or numbers like ستات *stAt* 'six', and هطعش *hTςš*

[2] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order)
أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
Â b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

‘eleven’. There are of course many open-class words that make YEMS different from MSA and other dialects. In particular, there are Turkish loanwords like بردق *brdq* ‘cup’ and ساني *sAny* ‘direct’.

4. Corpora

The MOR and YEMS corpora consist of sources of various genres, collected from both online and print materials, to cover all the aspects in these dialects. The MOR corpus has 64K words, while the YEMS corpus has 32.5K words; the genres and sources are summarized in Table 1 and Table 2 for MOR and YEMS, respectively.

The data collected from the Internet was all written in Arabic characters using spontaneous orthography. We transcribed the Roman alphabet sentences from the textbooks into the Arabic alphabet using Conventional Orthography for Dialectal Arabic (CODA).

4.1. MOR Corpus

Because much of the data comes from the Internet, there is some amount of code-switching between MOR and MSA, and sometimes French. The boundary between MOR and MSA is not always clear, especially for nouns and adjectives. In the corpus, tokens that exhibit MSA-only morphology are marked as MSA, so that they can be excluded from the morphological analyzer (e.g. هذين *haðayni* ‘these [m.dual]’). Words that might seem like MSA but nonetheless exhibit Moroccan morphology, such as technical terms, are considered dialectal.

The material in the various genres spans a wide range of content and registers. Comments on the Moroccan news website `hespress.com` have to do with issues such as sports, cinema, and education policy. The material from the forums includes 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.

4.2. YEMS Corpus

The social texts are taken from a Sanaani Radio Station program called *"msςd w msςdħ"* , the program addressed social issues and problems of the community. The oral interview transcripts are taken from the Semitisches Tonarchiv (see Section 2.). The interviews describe daily life, history and lifestyle in Sana’a. The folktales come from internet forums; they describe traditional stories handed down in Sana’a.

The wisdom and tales come from internet websites. These collected texts are a summary of the Wisdom and the Tales of the famous wise-man of Yemen *"ςly wld zAyd"*, a traditional oral tale passed through generations. Other texts are taken from social media, and include political events in Yemen, Sanaani jokes, religious sermons and transcripts that clarify the Sanaani dialect in MSA. The corpus shows distinctive sentences and phrases such as the sentence عنضوي نشرك شركة شرق عيغلقوا *ςnDwy nšrk šrkħšrq ςyγlqwA* ‘we will go to purchase meat before the store closes’, which contains words that distinguish Sanaani dialect from the other Yemeni dialects.

Genre	Source	# Tokens
Internet comments	`hespress.com`	8,909
Forums	`anaqamaghribia.com`	2,554
	`esrar.7olm.org`	2,509
Oral interviews	Appendix of (Brustad, 2000)	1,172
	Humans of Morocco	7,012
Folktales	`maghress.com`	5,704
Textbook examples	(The Peace Corps, 2011)	2,585
	(Chekayri, 2011)	4,858
Blog posts	`twishiat.com`	9,525
	`mysite.ma`	621
Humor	Facebook posts	16,756
	`as7apcool.com`	1,965
Total		64,170

Table 1: Sources of the Moroccan Arabic Corpus (MOR)

Genre	Source	# Tokens
Oral interviews	`Heidelberg uni.`	15,124
Social texts	`SanaaRadio`	5,515
Wisdoms and tales	`ye1.org`	1,500
Sanaani folktales	`ye1.org`	3,339
Sermons	Facebook posts	699
Poems	`g11y.com/vb`	906
Humor	Facebook posts	3,332
Explanation	`n-shbab.com`	1,704
Politic text	`marebpress.net`	326
Total		32,445

Table 2: Sources of Sanaani Yemeni Corpus (YEMS)

Each corpus is divided into three parts: DEV (roughly 10%), TRAIN (roughly 80%), and TEST (roughly 10%). Each part contains material from each genre, though not every source is divided into DEV, TRAIN, and TEST since some sources are quite small. Moreover, no document is split across multiple parts of the corpus. For example, there is no story or post where the beginning is in TRAIN and the end is in TEST. This ensures that a system is never trained and tested on data from the same document.

5. Annotation

The corpora are annotated using the DIWAN interface (Al-Shargi and Rambow, 2015). DIWAN assists a human annotator in annotating each token with morphological and semantic information, including the following fields:

- **Diac** is the token with spelling adjusted to be conform with the CODA guidelines, which specify a consistent orthographic system for writing Arabic dialects (Habash et al., 2012a). We created specific CODA guidelines for MOR and YEMS. However, despite the feature name, we do not use diacritics for MOR and YEMS.

- **Lex** is the lemma, or the citation form, of the token. For example, the lemma of وصحابه *wSHAbuh* ‘and his friends’ is صاحب *SAHb* ‘friend’.
- **BWhash** is the word broken down into prefixes, a stem, and suffixes, with each morpheme annotated with part of speech (POS) and other morphological information. The stem is marked by the symbol # on either side.
- **Gloss** is the English gloss of the word.
- There are features indicating proclitics and enclitics. The clitics are assigned slots: prc3, prc2, prc1, and prc0 for proclitics, and enc0, enc1, and enc2 for enclitics. A lower index indicates closer proximity to the stem.
- There are features indicating part of speech (POS), functional number and gender, and aspect. Functional number and gender refer to a word’s function, rather than its form. For example طلبة *Talabaħ* ‘students’ is functionally masculine plural, even though it ends in ة *ħ*, which is formally feminine singular.

For example, in the MOR sentence راه برّاف دناس ما فاهمينش *rAh bz∼Af dnAs mA fAhmiynš* ‘A lot of people really don’t understand’, the word فاهمينش*fAhmiynš* ‘understand-NEG’ gets the following annotation (which we show in transliteration for convenience; the annotation happens with the Arabic alphabet).

- **Diac**: fAhmynš
- **Lex**: fAhm
- **Bwhash**: +#fAhm/ACT_PARTIC# +yn/NSUFF_MASC_PL +š/NEG_PART
- **Gloss**: understanding
- **Clitics**: enc2:part_neg (‘enc2’ refers to the second slot for an enclitic. The other clitic slots are empty for this word.)
- **Other features**:
  - *part of speech*: active participle
  - *formal gender*: masculine
  - *functional gender*: neutral
  - *formal number*: plural
  - *functional number*: plural

The Moroccan corpus is currently 40.3% annotated. The annotation effort is ongoing.

The annotations follow the same format for the YEMS corpus. For example, in the sentence عنضوي نشرك شركة شرق عيغلقوا *ςnDwy nšrk šrkħšrq ςyγlqwA* ‘we will go to purchase meat before the store closes’, the word عيغلقوا *ςyγlqwA* ‘they will lock’ gets the following annotation:

- **Diac**: EyglqwA
- **Lex**: galaq
- **Bwhash**: +E/FUT_PART+y/IV3MP+#glq/IV# +wA/IVSUFF_SUBJ:3MP
- **Gloss**: lock
- **Clitics**: prc1:E_fut (‘prc1’ refers to the first slot for a proclitic. The other clitic slots are empty for this word)
- **Other features**:
  - *part of speech*: verb
  - *formal gender*: masculine
  - *functional gender*: masculine
  - *formal number*: plural
  - *functional number*: plural

The YEMS corpus has been annotated to 75%. The annotation effort for YEMS is also ongoing.

6. Morphological Analyzer

Next, we create two ALMOR databases (Habash, 2007) that represent the morphological analyzers for MOR and YEMS. The analyzers are constructed first by building complete inflectional classes (ICs) based on the corpus annotations. The construction of the ICs follows the technique we presented in (Eskander et al., 2013), where the ICs have all the possible morphosyntactic feature combinations for every lemma in TRAIN. Moreover, we extend the work presented in (Eskander et al., 2013) to cover any POS type, whether with clitics or without, in order to obtain rich morphological analyzers.

First, the entries in TRAIN are converted into paradigms, where each paradigm lists all the inflections of all morphosyntactic feature combinations for a specific lemma. The paradigms are then converted into inflectional classes (ICs), where stem entries are abstracted as templates by extracting out the root letters. We use the SCHLR template we defined in (Eskander et al., 2013), in which all long vowels, diacritics and hamzated letters remain in the template, and anything else is part of the root. (Note that this does not always result in the traditional notion of “root”, but rather in an operational notion that is useful for our purposes.) The generated ICs are then merged together into a smaller number of more condensed ICs, where two ICs merge if they share the same inflectional behavior. The ICs are then completed by exchanging affix and stem information among each other.

Table 3 shows the initial IC of the lemma زاد *zAd* ‘increase’ in YEMS. (We again show the entires in transliteration for convenience, though the system uses the Arabic alphabet.) The IC has initially three entries, which represent all the seen entries that correspond to the lemma زاد *zAd* ‘increase’ in TRAIN. The ICs are then completed for all morphosyntactic feature combinations. A portion of the completed IC of the lemma زاد *zAd* ‘increase’ is shown in Table 4.

Next, we use the completed ICs to build the MOR and YEMS morphological analyzers in the form of an ALMOR database (Habash, 2007), where the prefixes and suffixes are read directly from the ICs, while the stems are constructed by plugging the roots associated with the ICs into the stem templates in the ICs. We then construct three compatibility tables; prefix-stem, stem-suffix and prefix-suffix, based on the co-occurrence of the prefixes, stems and suffixes in the completed ICs.

We then extend the analyzers to allow for the recognition of spontaneous orthography that is not in CODA. We convert such input into a CODA-compliant form. This is done based on the orthographic transformations seen in the corpora between the raw input and the annotated data.

Tables 5 and 6 list the evaluation results of the MOR analyzer ($\text{ALMOR}_{\text{MOR}}$) on DEV and TEST, respectively,

Lemma: زاد *zAd*				
Features	Word	Initial IC		
		Prefixes	Stem	Suffixes
I1S	A+zyd	A	□y□	
P3MS	zAd		□A□	
P3MP	zAd+wA		□A□	wA

Table 3: The initial IC of the lemma زاد *zAd* 'increase' in YEMS. The first column lists the morphosyntactic features, while the second column lists the corresponding words segmented into prefixes+stem+suffixes. The third column shows the IC forms after stem abstraction.

Lemma: زاد *zAd*				
Features	Completed IC			Word
	Prefixes	Stem	Suffixes	
I1S	A	□y□		A+zyd
P3MS		□A□		zAd
P3MP		□A□	wA	zAd+wA
C2MS		□y□		zyd
I2FS	t	□y□	y	t+zyd+y
PART:$+I1S	$A	□y□		$A+zyd
CONJ:w+P3MP	w	□A□	wA	w+zAd+wA
CONJ:w+I1P	wn	□y□		wn+zyd
NEG:mA+P3MS+NEG:$	mA_	□A□	$	mA_+zAd+$
CONJ:w+P3MS+DO:3MS	w	□A□	h	w+zAd+h

Table 4: A portion of the completed IC of the lemma زاد *zAd* 'increase' in YEMS. The first column lists the morphosyntactic features, while the second column represents the abstracted forms of the completed IC. The third column lists the corresponding words segmented into prefixes+stem+suffixes after plugging the root "zd" into the stem templates.

while tables 7 and 8 list the evaluation results of the YEMS analyzer (ALMOR$_{\text{YEMS}}$) on DEV and TEST, respectively. (When evaluating on TEST, TRAIN and DEV are combined together and become the new TRAIN of the evaluated analyzer.) We compare the performance of the analyzers versus the SAMA$_{MSA}$ analyzer (Graff et al., 2009) (our MSA baseline), the ALMOR$_{\text{EGY}}$ analyzer (Habash et al., 2012b) (our EGY baseline) and a dialectal extended version of SAMA$_{MSA}$ (SAMA$_{ext}$) combined with ALMOR$_{\text{EGY}}$, which is the base for MADAMIRA, in addition to a simple lookup baseline (Lookup$_{\text{MOR}}$ for MOR and Lookup$_{\text{YEMS}}$ for YEMS). We also show the results when combining the different systems.

We use two main evaluation metrics to measure the performance of a morphological analyzer: 1) **Analyzer Token Recall**, which measures whether the hand-annotated analysis is automatically generated by the analyzer (usually, among other analyses), and 2) **OOV**, which represents the analyzer out-of-vocabulary cases.

The Analyzer Token Recall evaluates the following components:

- **POS** is the core POS tag of the word; a set of 36 tags that are used in MADA-ARZ.
- **POS5** is a reduced tag set of five tags based on traditional Arabic grammar.
- **Lemma** is the fully diacritized lemma.
- **CODA** is the undiacritized conventional spelling of the input word with normalized Alefs.
- **Stem** is the undiacritized stem of the word with normalized Alefs.
- **ALL** represents the conjunction of all five preceding metrics in one analysis.

7. Conclusion and Future Work

We have presented new corpora for Moroccan and Sanaani Yemeni dialectal Arabic. For both dialects, we have developed an orthographic convention for use in NLP. These corpora have been annotated morphologically. We have used the annotated corpora to train morphological analyzers and taggers.

To obtain the corpora and the morphological analyzers and taggers, please consult `http://volta.ccls.columbia.edu/~rambow/arabic-nlp/home.html`.

In future work, we will extend our approach to more Arabic dialects. We also intend to perform experiments to see whether we can leverage annotations from different dialects in training the morphological taggers.

8. Acknowledgments

Kaplan was supported by the National Science Foundation under an REU supplement to grant 1422863, Kathleen McKeown, PI. Portions of the annotations of the Yemeni and Moroccan corpora were funded by the Computational Approaches to Modeling Language (CAMeL) Lab at New York University Abu Dhabi.

	Analyzer Token Recall							
System	POS	POS5	Lemma	CODA	Stem	All	OOV	$\frac{Analyses}{Word}$
SAMA$_{MSA}$	88.2	97.8	86.0	95.8	92.3	79.9	17.1	9.5
ALMOR$_{\mathrm{EGY}}$	79.7	93.5	67.0	95.5	86.5	58.7	20.5	3.7
SAMA$_{ext}$+ALMOR$_{\mathrm{EGY}}$	90.0	98.3	87.6	96.4	94.3	81.7	11.8	15.9
Lookup$_{\mathrm{MOR}}$	62.3	80.0	57.2	92.5	69.7	55.2	45.6	0.6
ALMOR$_{\mathrm{MOR}}$	70.9	84.8	60.7	93.5	76.8	57.7	33.8	1.4
Lookup$_{\mathrm{MOR}}$+ALMOR$_{\mathrm{MOR}}$+SAMA$_{ext}$+ALMOR$_{\mathrm{EGY}}$	98.0	99.4	96.7	98.4	98.0	95.1	6.9	17.9

Table 5: MOR morphological analysis recall on DEV. The columns are described in Section 6.

	Analyzer Token Recall							
System	POS	POS5	Lemma	CODA	Stem	All	OOV	$\frac{Analyses}{Word}$
SAMA$_{MSA}$	94.3	99.0	94.7	97.9	95.8	90.4	16.2	9.1
ALMOR$_{\mathrm{EGY}}$	83.0	93.3	72.0	97.2	88.0	64.8	19.7	3.7
SAMA$_{ext}$+ALMOR$_{\mathrm{EGY}}$	95.0	99.1	95.2	98.1	97.2	91.0	10.8	15.7
Lookup$_{\mathrm{MOR}}$	62.8	78.4	55.4	94.7	70.7	54.0	46.2	0.7
ALMOR$_{\mathrm{MOR}}$	71.6	83.7	58.1	95.6	77.8	56.2	33.5	1.5
Lookup$_{\mathrm{MOR}}$+ALMOR$_{\mathrm{MOR}}$+SAMA$_{ext}$+ALMOR$_{\mathrm{EGY}}$	99.6	100.0	99.9	99.9	99.7	99.1	6.9	17.9

Table 6: MOR morphological analysis recall on TEST. The columns are described in Section 6.

9. Bibliographical References

Abdullah, A.-H. (1991). *Hawliat yamaniah: alyemen fi alqrn altasi' ashar almiladi*. dar Al-Hikma, Sana'a, Yemen.

Al-Iryani, M. A. (1996). *al-Mu'jam al-yemeni fi algah walturath*. dar alfkr, Damascus, Syria.

al Salam Al-Amri, A. (2000). *Texts in Sanani Arabic*. O. Harrassowitz, Wiesbaden, Germany.

Al-Shargi, F. and Rambow, O. (2015). Diwan: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, Beijing, China, July. Association for Computational Linguistics.

Behnstedt, P. (1985). *Die nordjemenitischen Dialekte*. L. Reichert, Wiesbaden, Germany.

Behnstedt, P. (2006). *Die nordjemenitischen Dialekte. II/3. Glossar fa - Ya*. L. Reichert, Wiesbaden, Germany.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multi-dialectal parallel corpus of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Boujelbane, R., Ellouze Khemekhem, M., and Belguith, L. H. (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, Nagoya, Japan.

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

Chekayri, A. (2011). *An Introduction to Moroccan Arabic and Culture*. Repertorio Español de Bibliografía Árabe e Islámica: 2010. Georgetown University Press.

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 *LREC*, pages 3782–3789.

Eskander, R., Habash, N., and Rambow, O. (2013). Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. In *Proceedings of tenth Conference on Empirical Methods in Natural Language Processing*.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Diab, M., and Rambow, O. (2012a). Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth Language Resources and Evaluation Conference*, pages 711–718, Istanbul, Turkey, May. European Language Resources Association.

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., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013). Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of NAACL-HLT*, Atlanta, GA.

Habash, N. (2007). Arabic Morphological Representations for Machine Translation. In Antal van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer.

Harrell, R., Sobelman, H., and Fox, T. (2004). *A Dictionary of Moroccan Arabic: Moroccan-English*. George-

	Analyzer Token Recall							
System	**POS**	**POS5**	**Lemma**	**CODA**	**Stem**	**All**	**OOV**	$\frac{Analyses}{Word}$
SAMA$_{MSA}$	86.5	96.1	81.5	94.1	89.4	75.1	14.9	10.7
ALMOR$_{\text{EGY}}$	81.3	91.9	69.6	93.5	86.3	60.5	19.3	4.4
SAMA$_{ext}$+ALMOR$_{\text{EGY}}$	91.1	97.4	86.7	94.6	93.5	80.5	8.7	19.3
Lookup$_{\text{YEMS}}$	72.5	83.6	71.2	93.2	77.9	68.6	41.9	0.9
ALMOR$_{\text{YEMS}}$	77.9	88.5	75.3	94.0	85.4	69.3	27.6	2.6
Lookup$_{\text{YEMS}}$+ALMOR$_{\text{YEMS}}$+SAMA$_{ext}$+ALMOR$_{\text{EGY}}$	97.1	99.2	95.7	96.6	97.6	92.4	5.4	22.8

Table 7: YEMS morphological analysis recall on DEV. The columns are described in Section 6.

	Analyzer Token Recall							
System	**POS**	**POS5**	**Lemma**	**CODA**	**Stem**	**All**	**OOV**	$\frac{Analyses}{Word}$
SAMA$_{MSA}$	89.8	96.8	83.7	95.5	89.6	77.8	12.7	11.7
ALMOR$_{\text{EGY}}$	83.0	92.5	69.1	94.5	86.7	60.6	17.5	4.8
SAMA$_{ext}$+ALMOR$_{\text{EGY}}$	93.9	97.9	89.3	95.8	93.5	83.0	7.3	21.0
Lookup$_{\text{YEMS}}$	71.4	82.6	69.3	93.3	75.7	67.6	43.3	1.0
ALMOR$_{\text{YEMS}}$	77.8	87.6	75.1	94.6	84.4	69.3	27.5	3.4
Lookup$_{\text{YEMS}}$+ALMOR$_{\text{YEMS}}$+SAMA$_{ext}$+ALMOR$_{\text{EGY}}$	98.3	99.3	97.4	97.6	97.9	95.1	4.6	25.4

Table 8: YEMS morphological analysis recall on TEST. The columns are described in Section 6.

town classics in Arabic language and linguistics. Georgetown University Press.

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

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, Doha, Qatar, October. Association for Computational Linguistics.

Jastrow, O. (1984). Zur Phonologie und Phonetik des Sana'nischen. In *Entwicklungsprozesse in der Arabischen Republik Jemen*, page 289âĂŞ304. L.Reichert, Wiesbaden, Germany.

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. LDC catalog number LDC99L22.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and using a pilot dialectal arabic treebank. In *LREC*, Genoa, Italy.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Naïm-Sanbar, S. (1994). Contribution à lâĂŹétude de lâĂŹaccent yéménite: Le parler des femmes de lâĂŹancienne génération. *Zeitschrift für Arabische Linguistik*, page 27.67âĂŞ89.

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

Salloum, W. and Habash, N. (2014). ADAM: Analyzer for Dialectal Arabic Morphology. *Journal of King Saud University-Computer and Information Sciences*, 26(4):372–378.

Smaïli, K., Abbas, M., Meftouh, K., and Harrat, S. (2014). Building resources for Algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*.

The Peace Corps. (2011). Moroccan Arabic. Based on *Moroccan Arabic* by Abdelghani Lamnaouar (1994); Abderrahmane Boujnab, editor.

Tratz, S., Briesch, D., Laoudi, J., Voss, C., and Holland, V. M. (2014). Language and dialect identification in social media analysis. *Proc. SPIE*, 9122:91220K–91220K–11.

Watson, J. C. (1993). *A syntax of San'ani Arabic*. O. Harrassowitz, Wiesbaden, Germany.

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