Building a Corpus for Palestinian Arabic: a Preliminary Study

Mustafa Jarrar, *Nizar Habash, Diyam Akra, Nasser Zalmout

Birzeit University, West Bank, Palestine
{mjarrar,nzalmout}@birzeit.edu, diyam@student.birzeit.edu

*New York University Abu Dhabi, United Arab Emirates
nizar.habash@nyu.edu

Abstract

This paper presents preliminary results in building an annotated corpus of the Palestinian Arabic dialect. The corpus consists of about 43K words, stemming from diverse resources. The paper discusses some linguistic facts about the Palestinian dialect, compared with the Modern Standard Arabic, especially in terms of morphological, orthographic, and lexical variations, and suggests some directions to resolve the challenges these differences pose to the annotation goal. Furthermore, we present two pilot studies that investigate whether existing tools for processing Modern Standard Arabic and Egyptian Arabic can be used to speed up the annotation process of our Palestinian Arabic corpus.

1. Introduction and Motivation

This paper presents preliminary results towards building a high-coverage well-annotated corpus of the Palestinian Arabic dialect (henceforth PAL), which is part of an ongoing project called *Curras*. Building such a PAL corpus is a first important step towards developing natural language processing (NLP) applications, for searching, retrieving, machine-translating, spell-checking PAL text, etc. The importance of processing and understanding such text is increasing due to the exponential growth of socially generated dialectal content at recent Social Media and Web 2.0 breakthroughs.

Most Arabic NLP tools and resources were developed to serve Modern Standard Arabic (MSA), which is the official written language in the Arab World. Using such tools to understand and process Arabic dialects (DAs) is a challenging task because of the phonological and morphological differences between DAs and MSA. In addition, there is no standard orthography for DAs. Moreover, DAs have limited standardized written resources, since most of the written dialectal content is the result of ad hoc and unstructured social conversations or commentary, in comparison to MSA's vast body of literary works.

The rest of this paper is structured as follows: We present important linguistic background in Section 2, followed by a survey of related work in Section 3. We then present the process of collecting the Curras Corpus (Section 4) and the challenges of annotating it (Section 5).

2. Linguistic Background

In this section we summarize some important linguistic facts about PAL that influence the decisions we made in this project. For more information on PAL and Levantine Arabic in general, see (Rice and Sa'id, 1960; Cowell, 1964; Bateson, 1967; Brustad, 2000; Halloun, 2000; Holes, 2004; Elihai, 2004). For a discussion of differences between Levantine and Egyptian Arabic (EGY), see Omar (1976).

2.1 Arabic and its dialects

The Arabic language is a collection of variants among which a standard variety (MSA) has a special status, while the rest are considered colloquial dialects (Bateson, 1967, Holes, 2004; Habash, 2010). MSA is the official written language of government, media and education in the Arab World, but it is not anyone's native language; the spoken dialects vary widely across the Arab World and are the true native varieties

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 18–27,
October 25, 2014, Doha, Qatar.

of Arabic, yet they have no standard orthography and are not taught in schools (Habash et al., 2012, Zribi et al., 2014).

PAL is the dialect spoken by Arabic speakers who live in or originate from the area of Historical Palestine. PAL is part of the South Levantine Arabic dialect subgroup (of which Jordanian Arabic is another dialect). PAL is historically the result of interaction between Syriac and Arabic and has been influenced by many other regional language such as Turkish, Persian, English and most recently Hebrew. The Palestinian refugee problem has led to additional mixing among different PAL sub-dialects as well as borrowing from other Arabic dialects. We discuss next some of the important distinguishing features of PAL in comparison to MSA as well as other Arabic dialects. We consider the following dimensions: phonology, morphology, and lexicon. Like other Arabic dialects, PAL has no standard orthography.

2.2 Phonology

PAL consists of several sub-dialects that generally vary in terms of phonology and lexicon preferences. Commonly identified sub-dialects include urban (which itself varies mostly phonologically among the major cities such as Jerusalem, Jaffa, Gaza, Nazareth, Nablus and Hebron), rural, and Bedouin. The Druze community has also some distinctive phonological features that set it apart. The variations are a miniature version of the variations in Levantine Arabic in general. Perhaps the most salient variation is the pronunciation of the /q/ phoneme (corresponding to MSA ق *q*[1]), which realizes as /'/ in most urban dialects, /k/ in rural dialects, and /g/ in Bedouin dialects. The Druze dialect retains the /q/ pronunciation. Another example is the /k/ phoneme (corresponding to MSA ك k), which realizes as /tš/ in rural dialects. These difference cause the word for قلب *qlb* 'heart' to be pronounced as /qalb/, /'alb/, /kalb/ and /galb/ and to be ambiguous out of context with the word كلب *klb* 'dog' /kalb/ and /tšalb/. And similarly to EGY (but unlike Tunisian Arabic), the MSA phoneme /θ/ (ث *θ*) becomes /s/ or /t/, and the MSA phoneme /ð/ (ذ *ð*) becomes /z/ or /d/ in different lexical contexts, e.g., MSA كذب *kðb* /kaðib/ 'lying' is pronounced /kizib/ in PAL and /kidb/ in EGY.

Similar to many other dialects, e.g. EGY and Tunisian (Habash et al., 2012; Zribi et al., 2014), the glottal stop phoneme that appears in many MSA words has disappeared in PAL: compare MSA رأس *rÂs* /ra's/ 'head' and بئر *bŷr* /bi'r/ 'well' with their Palestinian urban versions: /rās/ and /bīr/. Also, the MSA diphthongs /ay/ and /aw/ generally become /ē/ and /ō/; this transformation happens in EGY but not in other Levantine dialects such as Lebanese, e.g., MSA بيت *byt* /bayt/ 'house' becomes PAL /bēt/.

PAL also elides many short vowels that appear in the MSA cognates leading to heavier syllabic structure, e.g. MSA جبال /jibāl/ 'mountains' (and EGY /gibāl/) becomes PAL /jbāl/. Additionally long vowels in unstressed positions in some PAL sub-dialects shorten, a phenomenon shared with EGY but not MSA: e.g., compare /zāru/ (زاروا zAr+uwA) 'they visited' with /zarū/ (زاروه zAr+uw+h) 'they visited him'. Finally, PAL has commonly inserted epenthetic vowels (Herzallah, 1990), which are optional in some cases leading to multiple pronunciations of the same word, e.g., /kalb/ and /kalib/ (كلب *klb* 'dog'). This multiplicity is not shared with MSA, which has a simpler syllabic structure and more limited epenthesis than PAL.

2.3 Morphology

PAL, like MSA and its dialects and other Semitic languages, makes extensive use of templatic morphology in addition to a large set of affixations and clitics. There are however some important differences between MSA and PAL in terms of morphology. First, like many other dialects, PAL lost nominal case and verbal mood, which remain in MSA. Additionally, PAL in most of its sub-dialects collapses the feminine and masculine plurals and duals in verbs and

[1]Arabic orthographic transliterations are provided in the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007), *except where indicated*. HSB extends Buckwalter's transliteration scheme (Buckwalter, 2004) to increase its readability while maintaining the 1-to-1 correspondence with Arabic orthography as represented in standard encodings of Arabic, i.e., Unicode, etc. The following are the only differences from Buckwalter's scheme (indicated in parentheses): Ā آ (|), Â أ (>), ŵ ؤ (&), Ǎ إ (<), ŷ ئ (}), ħ ة (p), θ ث (v), ð ذ (*), š ش ($), Ď ظ (Z), ς ع (E), γ غ (g), ý ى (Y), ã ً (F), ũ ٌ (N), ĩ ٍ (K). Orthographic transliterations are presented in italics. For phonological transcriptions, we follow the common practice of using '/.../' to represent phonological sequences and we use HSB choices with some extensions instead of the International Phonetic Alphabet (IPA) to minimize the number of representations used, as was done by Habash (2010).

most nouns. Some specific inflections are ambiguous in PAL but not MSA, e.g., حبيت *Hbyt* /Habbēt/ 'I (or you [m.s.]) loved'.

Second, some specific morphemes are slightly or quite different in PAL from their MSA forms, e.g., the future marker is /sa/ in MSA but /Ha/ or /raH/ in PAL. Another prominent example is the feminine singular suffix morpheme (Ta Marbuta), which in MSA is pronounced as /at/ except at utterance final positions (where it is /a/). In some PAL urban sub dialects, it has multiple allomorphs that are phonologically and syntactically conditioned: /a/ (after non-front and emphatic consonants), /e/ (after front non-emphatic consonants), /it/ (nouns in construct state such as before possessive pronouns) and /ā/ (in deverbals before direct objects): e.g. بطة *bTħ* /baTT+a/ 'duck', حبة *Hbħ* /Habb+e/ 'pill', بطتنا *bTnA* /baTT+it+na/ 'our duck' and /mdars+ā +hum/ 'she taught them'.

Third, PAL has many clitics that do not exist in MSA, e.g., the progressive particle /b+/ (as in /b+tuktub/ 'she writes'), the demonstrative particle /ha+/ (as in /ha+l+bēt/ 'this house'), the negation cirmcumclitic /ma+ +š/ (as in /ma+katab+š/ 'he did not write') and the indirect object clitic (as in /ma+katab+l+ō+š/ 'he did not write to him'). All of these examples except for the demonstrative particle are used in EGY.

2.4 Lexicon

The PAL lexicon is primarily Arabic with numerous borrowings from many different languages. MSA cognates generally appear with some minor phonological changes as discussed above; a few cases include more complex changes, e.g. /biddi/ 'I want' is from MSA /bi+widd+i/ 'in my desire' or /illi/ 'relative pronoun which/who/that' which corresponds to a set of MSA forms that inflect for gender and number (الذي *Alðy*, التي *Alty*, etc.). Some common PAL words are portmanteaus of MSA words, e.g., /lēš / 'why?' corresponds to MSA /li+'ayy+i šay'/ 'for what thing?'. Examples of common words that are borrowed from other languages include the following:

- روزنامه /roznama/ 'calendar' (Persian)
- كندرة /kundara/ 'shoe' (Turkish)
- بندورة /banadora/ 'tomato' (Italian)
- بريك /brēk/ 'brake (car)' (English)
- تليفيزيون /talifizyon/ 'television' (French)
- محسوم /maHsūm/ 'checkpoint' (Hebrew)

3. Related Work

3.1 Corpus Collection and Annotation

There have been many contributions aiming to develop annotated Arabic language corpora, with the main objective of facilitating Arabic NLP applications. Notable contributions targeting MSA include the work of Maamouri and Cieri, (2002), Maamouri et al. (2004), Smrž and Hajič (2006), and Habash and Roth (2009). These efforts developed annotation guidelines for written MSA content producing large-scale Arabic Treebanks.

Contributions that are specific to DA include the development of a pilot Levantine Arabic Treebank (LATB) of Jordanian Arabic, which contained morphological and syntactic annotations of about 26,000 words (Maamouri et al., 2006). To speed up the process of creating the LATB, Maamouri et al. (2006) adapted MSA Treebank guidelines to DA and experimented with extensions to the Buckwalter Arabic Morphological Analyzers (Buckwalter, 2004). The LATB was used in the Johns Hopkins workshop on Parsing Arabic Dialect (Rambow et al., 2005; Chiang et al., 2006), which supplemented the LATB effort with an experimental Levantine-MSA dictionary. The LATB effort differs from the work presented here in two respects. First, the LATB corpus consists of conversational telephone speech transcripts, which eliminated the orthographic variations issues that we face in this paper. Secondly, when the LATB was created, there were no robust tools for morphological analysis of any dialects; this is not the case any more. We plan to exploit existing tools for EGY to help the annotation effort.

Other DA contributions include the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany, et al., 2002), which was developed as part of the CALLHOME Egyptian Arabic (CHE) corpus (Gadalla, et al., 1997). In addition to YADAC (Al-Sabbagh and Girju, 2012), which was based on dialectal content identification and web harvesting of blogs, micro blogs, and forums of EGY content. Similarly, the COLABA project (Diab et al., 2010) developed annotated dialectal content resources for Egyptian, Iraqi, Levantine, and Moroccan dialects, from online weblogs.

3.2 Dialectal Orthography

Due to the lack of standardized orthography guidelines for DA, along with the phonological differences in comparison to MSA, and dialectal variations within the dialects themselves, there are many orthographic variations for written DA content. Writers in DA, regardless of the context, are often inconsistent with others and even with themselves when it comes to the written form of a dialect; writing with MSA driven orthography, or writing words phonologically sometimes. These orthography variations make it difficult for computational models to properly identify and reason about the words of a given dialect (Habash et al, 2012a), hence, a conventional form for the orthographic notations is important. Within this scope, we can view this problem for Levantine dialects as an extension of the work of Habash et al. (2012a) who proposed the so-called CODA (Conventional Orthography for Dialectal Arabic). CODA is designed for the purpose of developing conventional computational models of Arabic dialects in general. Habash et al. (2012a) provides a detailed description of CODA guidelines as applied to EGY. Eskander et al. (2013) identify five goals for CODA: (i) CODA is an internally consistent and coherent convention for writing DA; (ii) CODA is created for computational purposes; (iii) CODA uses the Arabic script; (iv) CODA is intended as a unified framework for writing all DAs; and (v) CODA aims to strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities. CODA guidelines will be extended to cover PAL in this paper, as discussed in Section 5.3.

3.3 Dialectal Morphological Annotation

Most of the work that explored morphology in Arabic focused on MSA (Al-Sughaiyer and Al-Kharashi, 2004; Buckwalter, 2004; Habash and Rambow, 2005; Graff et al., 2009; Habash, 2010). The contributions for DA morphology analysis, however, are relatively scarce and are usually based on either extending available MSA tools to tackle DA specificities, as in the work of (Abo Bakr et al., 2008; Salloum and Habash, 2011), or modeling DAs directly, without relying on existing MSA contributions (Habash and Rambow, 2006). Due to the variations between MSA and DAs, available MSA tools and resources cannot be easily extended or transferred to work properly for DA (Maamouri, et al., 2006; Habash, et al., 2012b). Therefore, it is important to develop annotated and morpheme-segmented resources, along with morphological analysis tools, that are specific and tailored for DAs. One of the notable recent contributions for EGY morphological analysis was CALIMA (Habash et al., 2012b). The CALIMA analyzer for EGY and the commonly used SAMA analyzer for MSA (Graff et al., 2009) are central in the functioning of the EGY morphological tagger MADA-ARZ (Habash et al., 2013), and its successor MADAMIRA (Pasha et al., 2014), which supports both MSA and EGY.

The work we present in this paper builds on the shoulders of these previous efforts from the development of guidelines for orthography and morphology (in MSA and EGY) to the use of existing tools (specifically MADAMIRA MSA and EGY) to speed up the annotation process.

4. Corpus Collection

Written dialects in general tend to have scarce resources in terms of written literature; written materials usually involve informal conversations or traditional folk literature (stories, songs, etc.). It is therefore often difficult to find resources for written dialectal content. In addition, resources of dialectal content are prone to significant noise and inconsistency because they tend to lack standard orthographies and rely on ad hoc transcriptions and orthographic borrowing from the standard variety. In the case of Arabic, unlike MSA that dominates the formal and written content outlets, as in the press, scientific articles, books, and historical narration, DAs are more naturally used in traditional and informal contexts, such as conversations in TV series, movies, or on social media platforms, providing socially powered commentary on different domains and topics. And given the lack of standard orthography, there is common mixing of phonetic spelling and MSA-cognate-based spelling in addition to the so-called Arabizi spelling – writing DAs in Roman script, rather than Arabic script (Darwish, 2014 and Al-Badrashiny et al., 2014). Such noise imposes many challenges regarding the collection of high-coverage high-accuracy DA corpora. It is therefore important to remark that although *bigger is better* when it comes to corpus size, we focus more in this first iteration of our PAL corpus on precision and variety rather than mere

size. That is, we tried not only to manually select and review the content of the corpus, but also to assure that we covered a variety of topics and contexts, localities and sub-dialects, including the social class and gender of the speakers and writers. This is because such aspects help us discover new language phenomena in the dialect as will be discussed in the next section.

Table 1 presents the resources that we manually collected to build the PAL Curras corpus. There are 133 social media threads (about 16k words) from blogs (e.g., مدونة عبد الحميد العاطي Abdelhameed Alaaty's blog), forums (e.g., شبكة الحوار الفلسطيني The Palestinian dialogue network), Twitter, and Facebook. The collection was done by reading many discussion threads and selecting the relevant ones to assure diversity and PAL representative content. Content that is heavily written in a mix of languages, or a mix of other dialects was excluded. In the same way, we also manually collected some PAL stories, and a list of PAL terms and their meanings, which reflect additional diversity of topics, contexts, and social classes. About half of our corpus comes from 41 episode scripts from the Palestinian TV show وطن ع وتر "Watan Aa Watar". Each episode discusses and provides satirical critiques regarding different topics of relevance to the Palestinian viewers about daily life issues. The show's importance stems from the fact that the actors use a variety of Palestinian local dialects, hence enriching the coverage of the corpus.

Table 1. The Curras Corpus Statistics

Document Type	Word Tokens	Word Types	Documents
Facebook	3,120	1,985	35 threads
Twitter	3,541	2,133	38 threads
Blogs	8,748	4,454	37 threads
Forums	1,092	798	33 threads
Palestinian Stories	2,407	1,422	6 stories
Palestinian Terms	759	556	1 doc
TV Show: وطن ع وتر *Watan Aa Watar*	23,423	8,459	41 episodes
Curras Total	**43,090**	**19,807**	**191**

5. Corpus Annotation Challenges

This section presents our approach to annotating the Curras corpus. We start with a specification of our annotation goals, followed by a discussion of our general approach. We then discuss in more details two important challenges that need to be addressed for annotation of a new dialectal corpus: orthography and morphology.

5.1 Annotation Specification

The words are annotated in context. As such, the same word may receive different annotations in different contexts. We define the annotation of a word as a tuple $<w, w_B, c, c_B, l, p_B, g, i>$ described as follow. (Examples of such annotations are illustrated in Table 5.):

- ***w*: Raw (Unicode)** The raw input word defined as a string of letters delimited by white space and punctuation. The word is represented in Arabic script (Unicode).
- ***w_B*: Raw (Buckwalter)** The same raw input word in the commonly used Buckwalter transliteration (Buckwalter, 2004).
- ***c*: CODA (Unicode)** The Conventional Orthography (Habash et al., 2012) version of the input word.
- ***c_B*: CODA (Buckwalter)** The Buckwalter transliteration of the CODA form.
- ***l*: Lemma** The lemma of the word in Buckwalter transliteration. The lemma is the citation form or dictionary entry that abstracts over all inflectional morphology (but not derivational morphology). The lemma is fully diacritized. We follow the definition of lemma used in BAMA (Buckwalter, 2004) and CALIMA-ARZ (Habash et al., 2012b).
- ***p_B*: Buckwalter POS** The Buckwalter full POS tag, which identifies all clitics and affixes and the stem and assigns each a sub-tag. This representation treats clitics as separate tokens and abstracts the orthographic rewrites they undergo when cliticized. See the handling of the l/PREP+Al/DET in word #6 in Table 5. This representation is used by the LDC in the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) and tools such as MADAMIRA (Pasha et al., 2014). It is a high granularity representation that allows researchers to easily go to coarser granularity POS (Diab 2007; Habash, 2010; Alkuhlani et al., 2013). The Buckwalter POS tag can be fully diacritized or undiacritized. Given the added complexity of producing diacritized text manually by annotators, we opted at this stage to only use undiacritized forms.

- ***g*: Gloss** The English gloss, an informal semantic denotation of the lemma. In Tables 3-5, we only use one English word for space limitations.
- ***i*: Analysis** A specification of the source of the annotation, e.g., ANNO is a human annotator, and MADA is the MADAMIRA system with some minor or no automatic post-processing. In Tables 3 and 4, which are produced automatically, the Analysis field is replaced with a status indicating how usable the automatic annotation is.

5.2 General Approach

To speed up the process of annotating our corpus, we made the following decisions. First, and quite obviously from the previous section, we made a conscious decision to follow on the footsteps of previous efforts for MSA and EGY annotation done at the Linguistic Data Consortium and Columbia's Arabic Modeling group in terms of guidelines for orthography conventionalization and morphological annotation. This allows us to exploit existing guidelines with only essential modification to accommodate PAL and produce annotations that are comparable to those done for MSA and EGY. This, we hope, will encourage research in dialectal adaptation techniques and will make our annotations more familiar and thus usable by the community.

Second, and closely related to the first point, we exploit existing tools to speed up the annotation process. In this paper, we specifically use the MADAMIRA tool (Pasha et al., 2014) for morphological analysis and disambiguation of MSA and EGY. Our choice of using this tool is motivated by the assumption that EGY/MSA and PAL share many orthographic and morphological features. This assumption was validated by pilot experiments, presented below, and which show most of the PAL annotations can be generated automatically. However, a manual step is then needed to verify every annotation, to correct errors and fill in gaps. The manual annotation has not been completed yet as of the writing of this paper submission.

Finally, we made one major simplification to the annotations to minimize the load on the human annotator: we do not produce diacritized morphological analyses in the Buckwalter POS tag. The reasons for this decision are the following: (i) full diacritization is a complex task that most Arabic speakers do not do and thus it requires a lot of training and precious attention to detail; (ii) MSA and EGY produce many morphemes and lexical items that are quite similar to PAL except in terms of the short vowels (compare the lemmas for word #5 in Tables 3, 4 and 5); (iii) PAL has many cases of multiple valid diacritizations as mentioned above. While we think a convention should be defined to explain the variation and model it, it is perhaps the topic of a future effort that is more focused on PAL phonology. We make an exception for the lemmas and diacritize them since lemmas are important in indicating the core meaning of the word. In case of different pronunciations of the lemma, we choose the shortest.

5.3 A Conventional Orthography for PAL

As explained in Section 2, PAL, like other Arabic dialects, does not have a standard orthography. Furthermore, there are numerous phonological, morphological and lexical differences between PAL and MSA that make the use of MSA spelling as is undesirable. PAL speakers who write in the dialect produce spontaneous inconsistent spellings that sometimes reflect the phonology of PAL, and other times the word's cognate relationship with MSA. For example, the word for 'heart' (MSA قلب qalb) has four spellings that correspond to four sub-dialectal pronunciations: قلب *qlb* /qalb/, ألب Âlb /'alb/, كلب *klb* /kalb/, and جلب *jlb* /galb/. Similarly, the common shortening of some long vowels (from MSA to PAL) leads to different orthographies as in قانون *qAnwn* 'law' (MSA /qānūn/), which can also be written with a shortened first vowel قنون *qnwn* /'anūn/ reflecting the PAL pronunciation. PAL also has some clitics that do not exist in MSA, which leads to different spellings, e.g. the PAL future particle ح *H* /Ha/ can be written attached to or separate from the verb that follows it. Even when a morpheme exists in MSA and PAL, it may have additional forms or pronunciations. One example is the definite article morpheme ال *Al* /il/ which has a non-MSA/non-EGY allomorph /li/ when attached to nominals with initial consonant clusters. As a result, a word like /li+blād/ 'the homeland/countries' can be spelled to reflect the morphology as البلاد *AlblAd* or the phonology لبلاد *lblAd*, with the latter being ambiguous with 'for countries' (in PAL /la+blād/). Finally, there are words in PAL that have no cognate in MSA and as such have no

clear obvious spelling to go with, e.g., the word /barDo/ 'additionally' is spontaneously written as برضو *brDw*, برضه *brDh* and برضة *brDẖ*.

This, of course, is not a unique PAL problem. Researchers working on NLP for EGY and Tunisian dialects developed CODA guidelines for them (Habash et al., 2012a; Zribi et al., 2014). These guidelines were by design intended to apply (or be easily extended) to all Arabic dialects, but were only demonstrated for two. Our challenge was to take these guidelines (specifically the EGY version) and extend them. There were three types of extensions. First, in terms of phonology-orthography, we added the letter ك k to the list of root letters to be spelled in the MSA cognate to cover the PAL rural sub-dialects that pronounces it as /tš/. Second, in terms of morphology, we added the non-EGY demonstrative proclitic ه *h+* and the conjunction proclitic ت *t+* 'so as to' to the list of clitics, e.g., بهالبيت *bhAlbyt* 'in this house' and تيشوف *tyšwf* 'so that he can see'. Finally, we extended the list of exceptional words to cover problematic PAL words. All of the basic CODA rules for EGY (and Tunisian) are kept the same.

Pilot Study (I): We conducted a small pilot study in annotating the CODA for PAL words. We considered 1,000 words from 77 tweets in Curras. The CODA version of each word was created in context. 15.9% of all words had a different CODA form from the input raw word form. 42% of these changes involve consonants (two-fifths of the cases), vowels (one-fifth of the cases) and the hamzated/bare forms of the letter Alif ا *A*. Examples of consonant change can be seen in Table 5 (words #4 and #10). An additional 29% word changes involve the spelling of specific morpheme. The most common change (over half of the time) was for the first person imperfect verbal prefix ا *A* when following the progressive particle ب *b*: بكتب *bktb* as opposed to باكتب *bAktb*. About 18% of the changed words experience a split or a merge (with splits happening five time more than merges). An example of a CODA split is seen in Table 5 (word #9). Finally, only about 8% of the changed words were PAL specific terms; and less than 7% involved a typo or speech effect elongation. These results are quite encouraging as they suggest the differences between CODA and spontaneously written PAL are not extensive. Further analysis is still needed of course.

In Tables 3 and 4 (column CODA), we show the results of using the MADAMIRA-MSA and MADAMIRA-EGY systems on a set of ten words, while Table 5 shows the manually selected or corrected CODA. MADAMIRA generates a CODA version (contextually) by default. We expect the EGY version to be more successful than the MSA version in producing the CODA for PAL given the shared presence of many morphemes in EGY and PAL. However, when we ran the same set of words through MADAMIRA-EGY, we encountered many errors in words, morphemes and spelling choices in PAL that are different from EGY, e.g., the raw word منحب *mnHb* 'we love' (CODA بنحب *bnHb*) is analyzed as the EGY ما نحب *mA nHb* 'we do not love'!

5.4 Morphological Annotation Process and Challenges

To study the value of using an existing morphological analyzer for MSA or EGY in creating PAL annotations, we conducted the following pilot study.

Pilot Study (II): We ran the words from a randomly selected episode of the PAL TV show "Watan Aa Watar" (460 words) through both MADAMIRA-MSA and MADAMIRA-EGY. We analyzed the output from both systems to determine its usability for PAL annotations. We consider all analyses that are correct for PAL annotation or usable via simple post processing (such as removing CASE endings on MSA words) to be correct (as in word #2 in Tables 3-5). Words that receive incorrect analyses or no analyses require manual modifications.

The results of this experiment are summarized in Table 2. Table 3 and 4 illustrate sample results for ten words and Table 5 includes the manually created results.[2]

Table 2. Accuracy of automatic annotation of PAL text

Statistics	MADAMIRA MSA	MADAMIRA EGY
No Analysis	17.78%	7.24%
Wrongly Analyzed	18.43%	14.75%
Correctly Analyzed	63.79%	78.01%

The No Analysis (NA) words in Tables 2, 3 and 4 refer to the words that the morphological analyzer couldn't recognize. This failure may be

[2] The examples in Tables 3-5 are presented in the Buckwalter transliteration (Buckwalter, 2004) to match the forms as they appear in the annotated corpus.

a result of missing lexical entry, specific PAL morphology or typos. As expected, MADAMIRA-MSA had 2.5 times the number of NA cases compared to MADAMIRA-EGY. Examples include dialectal lexical terms (word #7) or dialectal morphology (words # 1 and #9).

The wrongly analyzed words are words that were assigned incorrect POS tag *in context*. For example, word #3 in Tables 3 and 4 is the result of mis-analyzing the proclitic l- as the preposition 'for/to' as opposed to the non-CODA spelling of the definite article in PAL. The analysis provided by MADAMIRA-EGY is correct for other contexts than the one illustrated here. Another example is word #8, which is a Levantine specific term hardly used in EGY and not used at all in MSA. MADAMIRA-MSA has a higher proportion of wrongly analyzed words than MADAMIRA-EGY.

Overall MADAMIRA-EGY produced analyses that were either correct and ready to use for PAL or requiring some minor modifications such as adjusting the vowels on the lemmas (e.g., word #5) in one of every five words.

Table 3 Automatic annotations by the MADAMIRA-MSA system. Entries with Status NA had no analysis.

	Raw		CODA		Lemma	Buckwalter POS (Diacritized)	Gloss	Status
1	ابوكوا	AbwkwA						NA
2	الاكل	AlAkl	الأكل	Al>kl	>akol	Al/DET+>akol/NOUN+a/CASE_DEF_ACC	eating	Usable
3	لبنوك	lbnwk	لبنوك	lbnwk	banok	li/PREP+bunuwk/NOUN+K/CASE_INDEF_GEN	bank	Wrong
4	التاني	AltAny	التأني	Alt>ny	ta>an~iy	Al/DET+ta>an~iy/NOUN	prudence	Wrong
5	الحمار	AlHmAr	الحمار	AlHmAr	HimAr	Al/DET+HimAr/NOUN+u/CASE_DEF_NOM	donkey	Usable
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	li/PREP+Al/DET+rAtib/NOUN+i/CASE_DEF_GEN	salary	Usable
7	ايوة	Aywp						NA
8	بدها	bdhA	بدها	bdhA	bud~	bud~/NOUN+i/CASE_DEF_GEN+hA/POSS_PRON_3FS	escape	Wrong
9	بنردلك	bnrdlk						NA
10	هدول	hdwl						NA

Table 4 Automatic annotations by the MADAMIRA-EGY system. Entries with Status NA had no analysis.

	Raw		CODA		Lemma	Buckwalter POS (Diacritized)	Gloss	Status
1	ابوكوا	AbwkwA	ابوكو	Abwkw	Abuw	Abuw/NOUN+kuw/POSS_PRON_3MS	father	Correct
2	الاكل	AlAkl	الأكل	Al>kl	>akl	Al/DET+>akol/NOUN	eating	Correct
3	لبنوك	lbnwk	لبنوك	lbnwk	bank	li/PREP+bunuwk/NOUN	bank	Wrong
4	التاني	AltAny	التاني	AltAny	tAniy	Al/DET+tAniy/ADJ_NUM	second	Usable
5	الحمار	AlHmAr	الحمار	AlHmAr	HumAr	Al/DET+HumAr/NOUN	donkey	Usable
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	li/PREP+Al/DET+rAtib/NOUN	salary	Correct
7	ايوة	Aywp	أيوه	>ywh	>ayowah	>ayowah/INTERJ	yes	Correct
8	بدها	bdhA	بدها	bdhA	bud~	bud~/NOUN+hA/POSS_PRON_3FS	escape	Wrong
9	بنردلك	bnrdlk	بنرد_لك	bnrd_lk	rad~	bi/PROG_PART+nu/IV1P+rud~/IV+li/PREP+ak/PRON_2MS	answer	Usable
10	هدول	hdwl						NA

Table 5 Manual Annotations in Curras. Entries with Analysis MADA were automatically converted and validated by the annotator. Entries with Analysis ANNO required some modification of the MADAMIRA output or were created from scratch.

	Raw		CODA		Lemma	Buckwalter POS (Undiacritized)	Gloss	Analysis
1	ابوكوا	AbwkwA	ابوكو	Abwkw	Abuw	Abw/NOUN+kw/POSS_PRON_3MS	father	MADA
2	الاكل	AlAkl	الأكل	Al>kl	>akl	Al/DET+>kl/NOUN	eating	MADA
3	لبنوك	lbnwk	البنوك	Albnwk	bank	Al/DET+bnwk/NOUN	bank	ANNO
4	التاني	AltAny	الثاني	AlvAny	vAniy	Al/DET+vAny/ADJ_NUM	second	ANNO
5	الحمار	AlHmAr	الحمار	AlHmAr	HmAr	Al/DET+HmAr/NOUN	donkey	MADA
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	l/PREP+Al/DET+rAtb/NOUN	salary	MADA
7	ايوة	Aywp	أيوه	>ywh	>ayowah	>ywh/INTERJ	yes	MADA
8	بدها	bdhA	بدها	bdhA	bid~	bd/NOUN+hA/POSS_PRON_3FS	want	ANNO
9	بنردلك	bnrdlk	بنرد_لك	bnrd_lk	rad~	b/PROG_PART+n/IV1P+rd/IV+l/PREP+k/PRON_2MS	answer	MADA
10	هدول	hdwl	هذول	h*wl	ha*A	h*wl/DEM_PRON	these	ANNO

5 Conclusion and Future Work

We presented our preliminary results towards building an annotated corpus of the Palestinian Arabic dialect. The challenges and linguistic variations of the Palestinian dialect, compared with Modern Standard Arabic, were discussed especially in terms of morphology, orthography, and lexicon. We also discussed and showed the potential, and limitations, of using existing resources, especially MADAMIRA-EGY, to semi-automate and speed up the annotation process.

The paper has also pointed out several issues that need to be considered and researched further, especially the development of Palestinian-specific morphological annotation and CODA guidelines, a Palestinian lexicon, and the extension of MADAMIRA to analyze Palestinian text. Our corpus will be further extended to include more text, and all lexical annotations (i.e., Lemmas) will be linked with existing Arabic ontology resources such as the Arabic WordNet (Black et al., 2006). The corpus will be publicly available for research purposes.

Acknowledgement

This work is part of the ongoing project *Curras*, funded by the Palestinian Ministry of Higher Education, Scientific Research Council. Nizar Habash performed most of his work on this paper while he was in the Center for Computational Learning Systems at Columbia University.

References

H. Abo Bakr, K. Shaalan, and I. Ziedan. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In The 6th International Conference on Informatics and Systems, INFOS2008. Cairo University, 2008.

M. Al-Badrashiny, R. Eskander, N. Habash, and O. Rambow. Automatic Transliteration of Romanized Dialectal Arabic. CoNLL, 2014.

S. Alkuhlani, N. Habash and R. M. Roth. Automatic Morphological Enrichment of a Morphologically Underspecified Treebank. In Proc. of Conference of the North American Association for Computational Linguistics (NAACL), Atlanta, Georgia, 2013.

R. Al-Sabbagh and R. Girju. YADAC: Yet another dialectal Arabic corpus. In Proc. of the Language Resources and Evaluation Conference (LREC), pages 2882–2889, Istanbul, 2012.

M. C. Bateson. Arabic Language Handbook. Center for Applied Linguistics, Washington D.C., USA, 1967.

W. Black, Elkateb, S., & Vossen, P. (2006). Introducing the Arabic wordnet project. In In Proceedings of the third International WordNet Conference (GWC-06).

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

T. Buckwalter. Buckwalter Arabic morphological analyzer version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0, 2004.

D. Chiang, M. Diab, N. Habash, O. Rambow, and S. Shareef. Parsing Arabic Dialects. In Proceedings of the European Chapter of ACL (EACL), 2006.

M. W. Cowell. A Reference Grammar of Syrian Arabic. Georgetown University Press, 1964.

Kareem Darwish. Arabizi Detection and Conversion to Arabic. In the Arabic Natural Language Processing Workshop, EMNLP, Doha, Qatar, 2014.

M. Diab. Towards an Optimal POS tag set for Modern Standard Arabic Processing. In Proc. of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, 2007.

M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba. COLABA: Arabic Dialect Annotation and Processing. LREC Workshop on Semitic Language Processing, Malta, 2010.

Y. Elihai. The olive tree dictionary: a transliterated dictionary of conversational Eastern Arabic (Palestinian). Washington DC: Kidron Pub, 2004.

R. Eskander, N. Habash, O. Rambow, and N. Tomeh. Processing Spontaneous Orthography. In Proceedings NAACL-HLT, Atlanta, GA, 2013.

H. Gadalla, H. Kilany, H. Arram, A. Yacoub, A. El-Habashi, A. Shalaby, K. Karins, E. Rowson, R. MacIntyre, P. Kingsbury, D. Graff, and C. McLemore. CALLHOME Egyptian Arabic Transcripts. Linguistic Data Consortium, Catalog No.: LDC97T19, 1997.

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

N. Habash and O. Rambow. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In ACL, Ann Arbor, Michigan, 2005.

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

N. Habash and R. Roth. CATiB: The Columbia Arabic Treebank. In ACL, 2009.

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

N. Habash, M. Diab, and O. Rabmow. (2012a) Conventional Orthography for Dialectal Arabic. In Proc. of LREC, Istanbul, Turkey, 2012.

N. Habash, R. Eskander, and A. Hawwari. (2012b) A Morphological Analyzer for Egyptian Arabic. In Proc. of the Special Interest Group on Computational Morphology and Phonology, Montréal, Canada, 2012.

N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh. Morphological Analysis and Disambiguation for Dialectal Arabic. In Proc. of NAACL, Atlanta, Georgia, 2013.

M. Halloun. A Practical Dictionary of the Standard Dialect Spoken in Palestine. Bethlehem University, 2000.

R. Herzallah. Aspects of Palestinian Arabic Phonology: A Nonlinear Approach. Ph.D. thesis, Cornell University. Distributed as Working Papers of the Cornell Phonetics Laboratory No. 4, 1990.

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

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. Egyptian Colloquial Arabic Lexicon. Linguistic Data Consortium, Catalog No.: LDC99L22, 1999.

M. Maamouri, A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow, and D. Tabessi. Developing and using a pilot dialectal Arabic treebank. In Proc. of LREC, Genoa, Italy, 2006.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004.

M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash and R. Eskander. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In Proc. of LREC, Reykjavik, Iceland, 2014.

M. Maamouri, and C. Cieri. Resources for Arabic Natural Language Processing at the Linguistic Data Consortium. In Proc. of the International Symposium on Processing of Arabic. Faculté des Lettres, University of Manouba, Tunisia, 2002.

M. Omar. Levantine and Egyptian Arabic: Comparative Study. Foreign Service Institute. Basic Course Series, 1976.

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

O. Rambow, D. Chiang, M. Diab, N. Habash, R. Hwa, K. Sima'an, V. Lacey, R. Levy, C. Nichols, and S. Shareef. 2005. Parsing Arabic Dialects. Final Report, 2005 JHU Summer Workshop.

F. Rice and M. Sa'id. Eastern Arabic: an introduction to the spoken Arabic of Palestine, Syria and Lebanon. Beirut: Khayat's 1960.

W. Salloum and N. Habash. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In Proc. of the First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties, Edinburgh, Scotland, 2011.

O. Smrž and J. Hajič. The Other Arabic Treebank: Prague Dependencies and Functions. In Ali Farghaly, editor, Arabic Computational Linguistics. CSLI Publications, 2006.

I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze Khmekhem, L. Hadrich Belguith, and N. Habash. A Conventional Orthography for Tunisian Arabic. In Proc. of LREC, Reykjavik, Iceland, 2014.