Lang Resources & Evaluation
DOI 10.1007/s10579-016-9370-7

ORIGINAL PAPER

Curras: an annotated corpus for the Palestinian Arabic dialect

Mustafa Jarrar[1] · Nizar Habash[2] · Faeq Alrimawi[1] ·
Diyam Akra[1] · Nasser Zalmout[2]

Abstract In this article we present Curras, the first morphologically annotated corpus of the Palestinian Arabic dialect. Palestinian Arabic is one of the many primarily spoken dialects of the Arabic language. Arabic dialects are generally under-resourced compared to Modern Standard Arabic, the primarily written and official form of Arabic. We start in the article with a background description that situates Palestinian Arabic linguistically and historically and compares it to Modern Standard Arabic and Egyptian Arabic in terms of phonological, morphological, orthographic, and lexical variations. We then describe the methodology we developed to collect Palestinian Arabic text to guarantee a variety of representative domains and genres. We also discuss the annotation process we used, which extended previous efforts for annotation guideline development, and utilized existing automatic annotation solutions for Standard Arabic and Egyptian Arabic. The annotation guidelines and annotation meta-data are described in detail. The Curras Palestinian Arabic corpus consists of more than 56 K tokens, which are

✉ Faeq Alrimawi
falrimawi@birzeit.edu

Mustafa Jarrar
mjarrar@birzeit.edu

Nizar Habash
nizar.habash@nyu.edu

Diyam Akra
diyam@student.birzeit.edu

Nasser Zalmout
nasser.zalmout@nyu.edu

[1] Birzeit University, Birzeit, Palestine

[2] New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Published online: 08 December 2016

annotated with rich morphological and lexical features. The inter-annotator agreement results indicate a high degree of consistency.

Keywords Palestinian Arabic · Palestinian corpus · Arabic morphology · Conventional Orthography for Dialectal Arabic · Dialectal Arabic · Word annotation

1 Introduction

Arabic is the official language of 23 countries, and is spoken by more than 300 million people as a first and second language. Arabic has multiple forms: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA is traced back to the Arabic of sixth and seventh century (pre-Islamic poetry and the Qur'an) and extending beyond the fifteenth century which witnessed the eclipse of Arab political power (Holes 2004). MSA generally describes the modern form of Arabic used for formal communication including news, media, education, and literature. DA, by contrast, is the informal day-to-day communication form of Arabic. Dialects can vary by way of life (*urban*, *rural*, or *Bedouin*), region, religion, social class, education, gender, and other factors. DAs are often divided regionally into five categories: (1) Levantine which is the dialect used in Levantine countries (Palestine, Jordan, Syria, and Lebanon), (2) Egyptian dialect which is the dialect spoken in Egypt and Sudan, (3) Maghrebi which is the dialect spoken in western Arab countries (Tunisia, Algeria, Libya, and Morocco), (4) Iraqi which is used in Iraq, and (5) Gulf which is used in the Arabic Gulf region. Although DA and MSA have large overlaps, differences can be observed in terms of their morphology, phonology, and lexicon. DA has no standard orthography, and there are numerous common words used in each specific region that are not part of MSA. Moreover, there are differences between DAs themselves. For example, the MSA سيكتب *sayaktubu*[1] 'he will write' is هيكتب *hayiktib* in Egyptian, غيكتب *γayiktib* in Moroccan, حيكتب *Hayuktub* in Levantine, and رح يكتب *raH yiktib* in Iraqi. Furthermore, within each DA, there are a number of sub-dialects.

While dialects are primarily the form of Arabic used in informal spoken genres (conversations, meetings, interviews, etc.), DA written content has been rapidly increasing over the internet in the past few years, especially through social media portals, as in Facebook, Twitter, weblogs, and others. Unlike in the past, where

[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). Arabic is written from right to left and with optional diacritics that are mostly used to mark vowels. Examples are vowelized as needed.

Arabic content was only written in MSA, the new DA web content is massively increasing, and the need to consume it is receiving more attention nowadays. For example, translating DA content automatically is becoming particularly important in social media platforms (Salloum and Habash 2013; Zbib et al. 2012), and so is applying sentiment analysis to DA customer reviews (of hotels, products, etc.) (Abdul-Mageed et al. 2012; Abdul-Mageed and Diab 2014). Although there are many tools and resources available for processing MSA text, these tools perform poorly on DA. For example, the MSA-mode of the morphological analyzer MAGEAD has been reported to have only 60 % coverage of Levantine Arabic verb forms (Habash and Rambow 2006); similar results have been shown for Egyptian Arabic (Habash et al. 2012b, 2013, Salloum and Habash 2014). Moreover, the use of a slightly modified version of an MSA morphological analyzer failed to produce results that can facilitate the annotation of Levantine Arabic texts during the development of the pilot Levantine Arabic Treebank (Maamouri et al. 2006). Thus, there is a need to develop NLP tools that directly address dialects, or a mix of DA and MSA text. An essential stage before developing NLP tools is building a corpus for a certain dialect. This motivates the importance of developing DA resources and tools.

Research on processing dialectal content can take several directions, including the development of morphological analysis techniques and tools, such as the development of the CALIMA morphological analyzer (Habash et al. 2012b), and the MADAMIRA tool for morphological analysis and disambiguation (Pasha et al. 2014), as well as machine translation of DA content (Zbib et al. 2012; Salloum and Habash 2013). The importance of building DA corpora started to emerge recently (Al-Sabbagh and Girju 2012; Bouamor et al. 2014; Meftouh et al. 2015; Khalifa et al. 2016). More attention is also given to the problem of developing a common DA orthography (Habash et al. 2012a).

Our contributions in this paper can be summarized as the following:

- We collected the first balanced corpus of the Palestinian Arabic dialect (PAL), which we name *Curras* ‘notebook’.
- We extended the Conventional Orthography for DA (CODA) guidelines (Habash et al. 2012a), to include the specifics of PAL. Three types of extensions were made: (1) phonology-orthography, (2) morphology, and (3) list of exceptional words to cover unique PAL words.
- We manually annotated Curras with rich morphological and lexical attributes such as part-of-speech (POS), stem, prefix, suffix, lemma, and gloss.
- We evaluated the annotation using inter-annotator agreement and showed that our annotations are highly consistent.

Our corpus, raw text, annotations, experiment data and the gold reference, and the annotation guidelines are fully available online of searching and downloading.[2]

It is worth noting that our preliminary findings were presented in (Jarrar et al. 2014), which has been significantly revised and extended. In this article we present

[2] Curras Portal http://portal.sina.birzeit.edu/curras.

the completely annotated corpus for PAL. Moreover, we present quantitative and qualitative evaluation of our annotations, and detail the PAL CODA guidelines.

The rest of this article is organized as follows: Sect. 2 introduces PAL and compares it with MSA; followed by literature review in Sect. 3. Then we present PAL text collection for the corpus in Sect. 4. In Sect. 5, we describe the methodology used for annotating the corpus. We present our extension of CODA guidelines for PAL in Sect. 6. In Sect. 7, we discuss several cases related to the annotation process. In Sect. 8, we present the evaluation of the annotations' accuracy and the inter-annotator agreement. Finally, in Sect. 9 we conclude our work, and discuss future work.

2 Palestinian Arabic

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 languages such as Turkish, Persian, English and Hebrew to some extent. 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.

2.1 Phonology

PAL consists of several sub-dialects that generally vary in terms of phonology and lexicon preferences. Commonly identified sub-dialects include urban, rural, Bedouin, and Druze. Urban itself varies phonologically among the major cities such as Jerusalem, Jaffa, Haifa, Gaza, Nazareth, Nablus and Hebron. 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. One of the most salient variation is the pronunciation of the MSA /q/ phoneme (corresponding the letter ق *q*), which is pronounced 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 MSA /k/ phoneme (corresponding to the letter ك *k*), which is pronounced as /tš/ in rural dialects.

Similar to many other dialects, e.g., Egyptian, Tunisian and Algerian (Habash et al. 2012a; Zribi et al. 2014; Saadane and Habash 2015), 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 Egyptian Arabic (henceforth, 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 *jbAl* /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+wA*) ‘they visited’ with /zarū/ (زاروه *zAr+w+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.

2.2 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. However, there are 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 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 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 is pronounced in MSA 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 *bTtnA* /baTT+it+na/ ‘our duck’ and مدرساهم *mdrsAhm* /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.

2.3 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 بودي *bwdy* /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 MSAلأي شيء *lÂy šy'* /li+'ayy+i šay'/ 'for what thing?'. Examples of common words that are borrowed from other languages include the following:

- روزنامه *rwznAmh* /roznama/ 'calendar' (Persian)
- كندرة *kndrħ* /kundara/ 'shoe' (Turkish)
- بندورة *bndwrħ* /banadora/ 'tomato' (Italian)
- بريك *bryk* /brēk/ 'brake (car)' (English)
- تليفزيون *tlyfzywn* /talifizyon/ 'television' (French)
- محسوم *mHswm* /maHsūm/ 'checkpoint' (Hebrew)

We further discuss PAL specifics, PAL guidelines, and provide examples on the usage of these guidelines in Sect. 6.

3 Literature review

This section reviews some of the related work in the field of Arabic and Arabic dialect processing. We first discuss the work related to MSA and then the work related to DA. In each of these subsections, we discuss monolingual corpora, parallel corpora, annotated corpora, and morphological processing tools. In discussing morphological processing, we distinguish between *morphological analysis* (out-of-context) and *morphological disambiguation* (in-context; akin to *morphological tagging*). In a final subsection, we discuss the issue of DA orthography. For surveys on resources for automatic processing of Arabic and its dialects, see (Al-Sughaiyer and Al-Kharashi 2004; Habash 2010; Shoufan and Al-Ameri 2015).

3.1 Resources for MSA

Monolingual and parallel corpora Important monolingual text collections of MSA include the Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006), the International Corpus of Arabic (ICA) (Alansary et al. 2007), and the very large Arabic Gigaword corpus (Parker et al. 2011). There are also many parallel resources for MSA with other languages, including the United Nations corpus (Arabic, Chinese, English, French, Russian and Spanish) (Rafalovitch and Dale 2009), and parallel Arabic-English corpora generated under large DARPA programs such as GALE and BOLT (Olive et al. 2011).

Morphological analysis and disambiguation tools There has been a considerable amount of work done on developing morphological analyzers for MSA. One of the most commonly used MSA analyzers is the Standard Arabic Morphological Analyzer (SAMA) (Graff et al. 2009), which is based on, and updates the Buckwalter Arabic Morphological Analyzer (BAMA). Other contributions are presented in (Beesley 1996; Al-Sughaiyer and Al-Kharashi 2004; Buckwalter 2004; Attia 2006; Smrž 2007). In terms of morphological disambiguation of the most commonly used tools include MADA+TOKAN (Habash and Rambow 2005;

Habash et al. 2009); AMIRA (Diab et al. 2007) and the most recent system that combines these two efforts, MADAMIRA (Pasha et al. 2014).

Annotated corpora One of the most important and widely used annotated MSA resources is the Penn Arabic Treebank (PATB) (Maamouri et al. 2004), which was developed at the Linguistic Data Consortium (LDC). This treebank consists of morphologically and syntactically annotated data (mostly newswire). The PATB syntactic representation is phrase structure with morphological tags from BAMA (Buckwalter 2004). Smrž and Hajic (2006) developed the Prague Arabic Dependency Treebank (PADT), which used multi-level linguistic annotations based on the Functional Generative Description theory. Habash and Roth (2009) built the Columbia Arabic Treebank (CATiB) with an approach emphasizing faster production with less linguistic information. CATiB depended on two basic ideas to speed the production process which are: (1) avoid the annotation of redundant linguistic information, e.g., nominal case markers, and (2) the use of a linguistic representation and terminology inspired by Arabic's long tradition of syntactic studies.

3.2 Resources for DA

Monolingual corpora One of the earliest efforts on DA corpus creation is the CALLHOME Egyptian Arabic (CHE) (Gadalla et al. 1997), which consisted of transcripts of EGY phone conversations. The COLABA project (Diab et al. 2010) collected DA resources (mainly for Egyptian and Levantine) from harvesting online weblogs. Resource creation for COLABA was semi-automatic where DA queries are used to harvest DA data. This is not the case in PAL corpus where we manually collected resources to accomplish high degree of coverage and annotation accuracy. Zaidan and Callison-Burch (2011) crawled three Arabic Newspaper websites and extracted readers commentary and built the Arabic Online Commentary dataset. Yet Another Dialectal Arabic Corpus (YADAC) (Al-Sabbagh and Girju 2012) presented a multi-genre DA corpus with focus on Egyptian Arabic. YADAC is based on dialectal content identification and web harvesting of blogs, micro blogs, and forums of Egyptian content. Most recently, Khalifa et al. (2016) presented a large-scale corpus for Gulf Arabic containing over 100 million words.

Parallel corpora There are no naturally occurring parallel corpora for DAs. To manage the lack of parallel resources for DA, some researchers have investigated the approach of bridging via MSA: Salloum and Habash (2013) translated DA to MSA as a pivot to translate to English. Similar work was done by Sajjad et al. (2013) and by Sawaf (2010). Other researchers developed different approaches for machine translation from DA to English: Zbib et al. (2012) described a crowd-resource approach to translate from Levantine and Egyptian into English and created two bilingual corpora. Bouamor et al. (2014) and Meftouh et al. (2015) independently developed multi-dialectal parallel corpora. Other parallel resources included the work of Al-Sabbagh and Girju (2010), who created a DA-to-MSA lexicon by mining the web; and Tharwa (Diab et al. 2014), a large-scale three-way DA-MSA-English Lexicon, containing over 73,000 Egyptian Arabic entries.

Morphological analysis and disambiguation tools Contributions towards DA morphology analysis are relatively scarce and recent. Some contributions are based on extending existing MSA tools to DA (Bakr et al. 2008; Salloum and Habash 2011, 2014). However, extending and applying MSA morphological analyzers to DA is complex and does not provide a complete solution (Maamouri et al. 2006), as there are variations between MSA and DA. A supervised algorithm for online morpheme segmentation on DA was developed by Riesa and Yarowsky (2006) for DA-to-English machine translation. Another approach, is developing analyzers for DA directly without depending on MSA tools. For example, MAGEAD (Habash and Rambow 2006) is the first morphological analyzer and generator for an Arabic dialect that includes a root-and-pattern analysis. Another recent notable morphological analyzer is CALIMA (Habash et al. 2012b) which is the Columbia Arabic Language and dIalect Morphological Analyzer for Egyptian Arabic. CALIMA was built by extending the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al. 2002) to provides a linguistically accurate, large-scale morphological analyzer that follows the POS guidelines used in the LDC's Arabic treebanks. CALIMA and SAMA analyzers are used in MADA-ARZ (Habash et al. 2013), a Morphological Analysis and Disambiguation of Egyptian Arabic (ARZ). A successor of MADA-ARZ is MADAMIRA (Pasha et al. 2014) which provides support for both MSA and Egyptian Arabic.

Annotated corpora In the context of developing treebanks for DA, a pilot Levantine Arabic Treebank (LATB) of Jordanian Arabic was built by Maamouri et al. (2006) at the LDC. The LATB contains morphological and syntactic annotations of about 26,000 words of Levantine Arabic conversational telephone speech. The LATB annotation guidelines were based on the MSA PATB with extensions. The approach for creating the LATB corpus is different from our approach for creating the Curras PAL corpus in two ways: (1) the LATB corpus consists of conversational telephone speech transcripts, which eliminated the orthographic variations issue present in our corpus since we collected PAL texts from written resources, and (2) at the time of the development of the LATB corpus, there were no robust morphological analyzers for any dialect, which is not the case for our corpus where we are able to exploit existing tools to facilitate the annotation process. The Egyptian Arabic Treebank (Maamouri et al. 2014) was developed by collecting informal content, e.g., discussion form texts. The development of this Treebank was in parallel with the development of a morphological analyzer for Egyptian Arabic, CALIMA (Habash et al. 2012b), where there was feedback loop and synchronization between them. The annotation process of the Egyptian Arabic Treebank depended on early versions of CALIMA analyzer which had many holes, and annotators were allowed to manually annotate words, so such entries were prone to errors. In PAL corpus we take advantage of MADAMIRA (Pasha et al. 2014) disambiguation tool, which includes CALIMA, and the DIWAN annotation interface (Al-Shargi and Rambow 2015) to minimize errors resulting from manual annotations.

Overall, the reviewed literature shows that many contributions have been made to MSA, less towards DAs, and fewer still targeted Levantine Arabic.

3.3 Conventional orthography for DA

DA lacks the existence of standard orthography guidelines, while MSA has an established orthographic standard. Arabic speakers writing in DA are often inconsistent with each other, and even with themselves. They may write words in a way that reflects the words' phonology or the MSA cognate the DA words are related to. The phonological variations among dialects themselves lead to even greater orthographic variations in written DA content. In addition, the so-called Arabizi spelling, which is writing DA in Roman script rather than Arabic script (Darwish 2013), poses more challenges to the process of collecting and processing DA text. These DA orthographic variations pose many challenges for tools and computational models to process DA words (Habash et al. 2012a). Hence, consistent and coherent conventional orthography guidelines are required to overcome this issue. A Conventional Orthography for Dialectal Arabic (CODA) (Habash et al. 2012a) was proposed as a solution to this problem. CODA is designed to develop conventional computational models of Arabic dialects in general. CODA guidelines and principles are created with a main goal of making CODA as a common conventional orthography that can be used for all Arabic dialects. CODA guidelines were developed for EGY (Habash et al. 2012a), and there are ongoing extensions to other dialects, e.g., Tunisian and Algerian (Zribi et al. 2014; Saadane and Habash 2015). In Sect. 6 we discuss the specifics of PAL and its CODA guideline extensions. A system called CODAFY (Eskander et al. 2013) was proposed as a general orthographic preprocessor that automatically converts spontaneous orthography of written EGY into CODA orthography. The CALIMA analyzer (Habash et al. 2012b) and MADAMIRA disambiguation tool (Pasha et al. 2014) utilize CODA internally and include *CODAfication* as part of the analysis and disambiguation processes. Finally, attempts to convert Arabizi text into Arabic script following the CODA guidelines was demonstrated by Eskander et al. (2014), who achieved an accuracy of 83.8 % on this task.

4 Corpus collection

This section presents the process of collecting our Curras corpus' PAL raw text. Written resources for dialects in general are relatively scarce; unlike MSA that dominates the written resources, as in the news media, education, science, and books, DAs are used only in informal contexts, such as conversations in a TV series, movies, and recently written DA started to emerge in social media platforms (Facebook, Twitter, blogs, etc.). Although DA is used in informal contexts, DA conveys socially powered commentary on different domains and topics, from personal narratives to traditional folk literature (stories, songs, etc.). Finding and collecting resources for PAL was a very difficult task. The high degree of spelling variation resulting from lack of standard orthography, the existence of different sub-dialects, and common use of different writing scripts (Arabic vs. Arabizi) make DA resources prone to significant noise and inconsistency, which challenges techniques using query matching to identify dialectal text in the specific dialect of interest.

Table 1 Statistics about the collected resources

Resource	Tokens (%)	Documents
Facebook	4852 (8.6)	35 threads
Twitter	4694 (8.3)	38 threads
Blogs	11,245 (19.8)	37 threads
Forums	1027 (1.8)	33 threads
Palestinian stories	3149 (5.6)	6 stories
Palestinian terms	1468 (2.6)	1 doc
TV Shows: وطن ع وتر *Watan Aa Watar*	30,265 (53.4)	41 episodes
Total	56,700	190

In our Curras PAL corpus we focus on precision and variety rather than on mere size. Our approach to collecting PAL content is described as follows:

- Manually collect and review resources for PAL, then select and determine suitable PAL content. The selection and reviewing of resources was performed by native PAL speakers.
- Cover a variety of contexts, subjects, and sub-dialects, including the social class and gender of the speakers and writers. This is achieved through careful selection of content from diverse resources reflecting different PAL contexts.
- Ignore content that is heavily written in a mix of languages or a mix of other dialects. We also do not include Arabizi PAL text.

The corpus data was collected from variety of resources, as described below. Table 1 shows the number of tokens (including words, digits, and punctuations) collected for each of the resources.

- *Facebook* We investigated many Facebook pages and manually extracted PAL content from different Palestinian pages, e.g., the page of "يما بديش أتجوز" "Mom, I don't want to get married". We included status posts and comments that are written in PAL. This represents about 8.6 % of our corpus.
- *Twitter* We collected PAL content from tweets of different Palestinian accounts; Tweets represent about 8.3 % of the corpus.
- *Blogs* We obtained PAL content through analyzing a number of blogs, e.g., the blog of "عبد الحميد العاطي" "Abdelhameed Alaaty's". We manually selected appropriate content. Blogs represent 19.8 % of our corpus.
- *Forums* We investigated the forum "شبكة الحوار الفلسطيني" "The Palestinian Dialogue Network". We carefully selected relevant PAL content to assure corpus precision and diversity. This represents 1.8 % of our corpus.
- *Palestinian stories* We collected a number of stories written in PAL from different resources and forums, e.g., "قصة سندريلا" "The Story of Cinderella". Stories of this kind are about 5.6 % of our corpus.

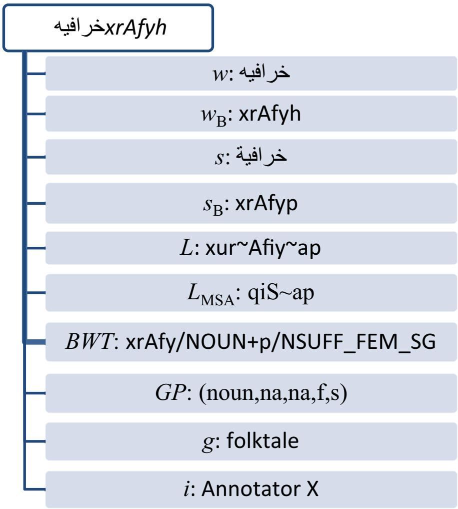


Fig. 1 Annotation of the word خرافيه *xrAfyh* ‘folktale’

- *Palestinian terms* We collected about 550 Palestinian terms and their meanings, e.g., “حاكورة = قطعة أرض صغيرة” “Hakura = a small piece of land”, from different websites which enriched our corpus. Terms represent 2.6 % of our corpus.
- *TV Shows* We were able to collect 41 episode scripts from a Palestinian TV show called “وطن ع وتر” “Watan Aa Watar”. The show discusses, and provides satirical critiques of, different topics of relevance to its 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. This TV show provides about 53.4 % of our corpus.

5 Corpus annotation

This section presents the approach we used to annotate our Curras PAL corpus. First we define the annotation metadata. Then, we describe the process and tools used for annotating each word in context.

5.1 Annotation definition and metadata

By *word annotation* we mean adding metadata to a word *in-context*, including its morphology, semantics, and other aspects. Figure 1 shows the word خرافيه *xrAfyh* ‘folktale’ as an example of information associated with a word after being annotated. We define an annotation as a tuple, of 10 tags:

$$\langle w, w_B, s, s_B, L, L_{MSA}, BWT, GP, g, i \rangle,$$

where:

1. ***w*: Raw Word (in 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).
2. ***w_B*: Raw Word (in Buckwalter),** the same raw input word ***w***, but in the Buckwalter transliteration (Buckwalter 2004; Habash et al. 2007).
3. ***s*: Surface Word (in Unicode),** the word ***w***, but written in the CODA Conventional Orthography (Habash et al. 2012a), and represented in Unicode. See Sect. 6 below.
4. ***s_B*: Surface Word (in Buckwalter),** the same as ***s***, but written in the Buckwalter transliteration.
5. ***L*: Lemma** (in Buckwalter) of ***w***. The lemma is the citation form or dictionary entry that abstracts over all inflectional morphology (but not derivational morphology). The lemma is fully diacritized. It is defined for verbs as the past, singular, masculine, 3rd person form of the verb; and for nouns is defined as the masculine singular form of the noun or the feminine singular if no masculine form exists (Buckwalter 2004; Habash et al. 2012b).
6. ***$L_{\mathbf{MSA}}$*: MSA Lemma** (in Buckwalter) of ***w***. This is similar to ***L***, but differs in that the value must be the MSA translation of the word. For example, the dialectal word يروح *yrwH* 'he goes' has the Lemma ***L*** راح *rAH* 'to go', and the MSA Lemma ***$L_{\mathbf{MSA}}$*** ذَهَب *ðahab (*ahab in Buckwalter transliteration)*.
7. ***BWT: Buckwalter POS Tag*** (in Buckwalter): the full POS tag, which specifies all clitics, affixes and the stem, and assigns each a subtag. This representation treats clitics as separate tokens and abstracts the orthographic rewrites they undergo when cliticized. This representation is used in the BAMA, SAMA and CALIMA analyzers, as well as the PATB and MADAMIRA. The Buckwalter POS tags 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. See Habash (2010) for a detailed description of this POS tag set and others.
8. ***GP:* Grammatical Properties:** G*P* consists of the following five feature-values pairs:
   - ***POS:* Part-of-Speech** of *w* is a coarse POS that does not specify any inflectional features, e.g., verb or noun. This POS feature corresponds to the **pos** feature in MADAMIRA and has 32 values (See Table 3).
   - ***P*: Person** is the grammatical person of the verb; it can take one of four values: **1** for first person, **2** for second person, **3** for third person, and **na** (not applicable) if the word is not a verb.
   - ***A*: Aspect** represents the time of the verb. It takes four values: **i** for imperfective verbs, **p** for perfective verbs, and **c** for command verbs, and **na** if the word is not a verb.
   - ***G*: Gender** specifies the grammatical gender of the word. It takes three values: **m** for masculine, **f** for feminine, and **na** in case this attribute is not applicable to the word.

- ***N*: Number** represents the grammatical number of the word. It takes four values: **s** for singular, **d** for dual, **p** for plural, and **na** when this attribute is not applicable.

  The gender and number features represent what is called functional morphology, which may be different from the gender and number form-based morphology expressed in BWT. This captures phenomena such as the irregular (*broken*) plurals in Arabic where the word may have a singular affix but is actually plural, e.g., التوانسة *AltwAnsħ* 'the Tunisians' ends with the *ħ* feminine singular morpheme but is a masculine plural word. For more information on the issues of form and function in Arabic morphology, see (Smrž 2007; Alkuhlani and Habash 2011).

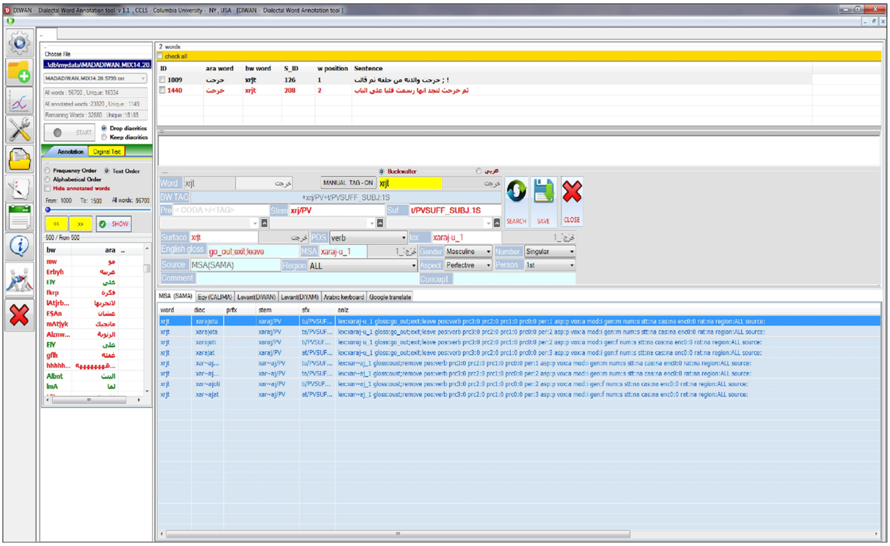
9. ***g*: Gloss** is the English gloss, an informal semantic denotation of the lemma.
10. ***i*: Analysis** specifies of the source of the annotation. This could be the annotator name, or the tool/database, e.g., SAMA, CALIMA, MADAMIRA, or combination of annotator name and SAMA, CALIMA, or MADAMIRA.

5.2 The annotation process and methodology

To facilitate and speed up the process of annotating a word, we made the following decisions. First, we made a conscious decision to follow on the footsteps of previous efforts for MSA and EGY annotation done at the LDC and Columbia's Arabic Dialect Modeling (CADiM) group in terms of guidelines for orthography conventionalization and morphological annotation. This allows us to exploit existing guidelines with only essential modifications 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, we exploited existing tools to speed up the annotation process. We specifically used the MADAMIRA tool (Pasha et al. 2014) for morphological analysis and disambiguation of MSA and EGY to create our initial annotations. Our choice for using this tool is motivated by the assumption that MSA, EGY and PAL share many orthographic and morphological features. This assumption was validated by pilot experiments presented in (Jarrar et al. 2014) which showed that MADAMIRA EGY returns correct analysis for about 78 % of the text.

Third, to further facilitate the annotation process, we used the DIWAN Dialect Word Annotation tool (Al-Shargi and Rambow 2015). We collaborated actively with the developers of DIWAN to extend the tool during the annotation of our corpus, by providing constant feedback and making feature requests. DIWAN provides a graphical user interface (see Fig. 2) for annotators to help in the manual annotation process. At its core, DIWAN uses MADAMIRA as the morphological analyzer and disambiguation tool for MSA and EGY. The user interface displays information returned from MADAMIRA. DIWAN displays the different types of metadata to the annotators, such as prefix, stem, suffix, lemma, gloss, etc., in drop down lists, and allows the annotators to edit these fields. In addition, it allows the annotators to search for and modify the analysis of a given word, from existing



DIWAN - Dialectal Word Annotation tool v 1.1 , CCLS - Columbia University - NY , USA - [DIWAN - Dialectal Word Annotation tool]

Choose File
..\db\mydata\MADADIWAN.MIX14.20.
MADADIWAN.MIX14.20.5799.txt
All words : 56700 , Unique: 16334
All annotated words :23820 , Unique : 1149
Remaining Words : 32880 , Unique :15185
START ● Drop diacritics ○ Keep diacritics
Annotation | Orginal Text
○ Frequency Order ● Text Order
○ Alphabetical Order
□ Hide annotated words
From: 1000 To: 1500 All words: 56700
<< >> SHOW
500 / From 500

bw	ara
mw	مو
Erbyh	عربيه
ElY	على
fkrp	فكرة
lAtjrb...	لاتجربها
E$An	عشان
mAtjyk	ماتجيك
Alznw...	الزنوبة
ElY	على
gflh	غفله
hhhhh...	هههههههه...
Albnt	البنت
lmA	لما

2 words
□ check all

ID	ara word	bw word	S_ID	w position	Sentence
□ 1009	خرجت	xrjt	126	1	! ; خرجت والدنه من خلغه نم قالت
□ 1440	خرجت	xrjt	208	2	ثم خرجت لنجد انها رسمت قلبا على الباب

● Buckwalter ○ عربي
Word: xrjt خرجت | MANUAL TAG - ON: xrjt خرجت
BW TAG: +xrj/PV+t/PVSUFF_SUBJ:1S
Pre: < CODA >/<TAG> | Stem: xrj/PV | Suf: t/PVSUFF_SUBJ:1S
Surface: xrjt خرجت | POS: verb | lex: xaraj-u_1 خَرَج_1
English gloss: go_out;exit;leave | MSA: xaraj-u_1 خَرَج_1 | Gender: Masculine | Number: Singular
Source: MSA(SAMA) | Region: ALL | Aspect: Perfective | Person: 1st
Comment: | Concept:

MSA (SAMA) | Egy (CALIMA) | Levant(DIWAN) | Levant(DIYAM) | Arabic keyboard | Google translate

word	diac	prfx	stem	sfx	anlz
xrjt	xarajotu		xaraj/PV	tu/PVSUF...	lex:xaraj-u_1 gloss:go_out;exit;leave pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:1 asp:p vox:a mod:i gen:m num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xarajota		xaraj/PV	ta/PVSUF...	lex:xaraj-u_1 gloss:go_out;exit;leave pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:2 asp:p vox:a mod:i gen:m num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xarajoti		xaraj/PV	ti/PVSUF...	lex:xaraj-u_1 gloss:go_out;exit;leave pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:2 asp:p vox:a mod:i gen:f num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xarajat		xaraj/PV	at/PVSUF...	lex:xaraj-u_1 gloss:go_out;exit;leave pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:3 asp:p vox:a mod:i gen:f num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xar~aj...		xar~aj/PV	tu/PVSUF...	lex:xar~aj_1 gloss:oust;remove pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:1 asp:p vox:a mod:i gen:m num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xar~aj...		xar~aj/PV	ta/PVSUF...	lex:xar~aj_1 gloss:oust;remove pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:2 asp:p vox:a mod:i gen:m num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xar~ajoti		xar~aj/PV	ti/PVSUF...	lex:xar~aj_1 gloss:oust;remove pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:2 asp:p vox:a mod:i gen:f num:s stt:na cas:na enc0:0 rat:na region:ALL source:
xrjt	xar~ajat		xar~aj/PV	at/PVSUF...	lex:xar~aj_1 gloss:oust;remove pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:3 asp:p vox:a mod:i gen:f num:s stt:na cas:na enc0:0 rat:na region:ALL source:

Fig. 2 DIWAN tool GUI

MSA and EGY annotations. We conducted an initial pilot annotation before introducing DIWAN in our pipeline. Using DIWAN did not only help in speeding up the annotation, but also in significantly reducing the number of errors and typing mistakes that our annotators produced. We discuss some of the annotation techniques we used in DIWAN in Sect. 7.

Two annotators (A1 and A2) annotated the corpus over a period of 1-year, part time. A portion of the corpus was doubly annotated to allow us to measure inter-annotator agreement (see Sect. 8).

5.3 Corpus statistics

Out of the 56,700 token instances (16,416 unique types) collected, we annotated 55,960 tokens (98.7 %), corresponding to 16,018 (97.6 %) types. The rest of the tokens were not annotated since they contained a lot of typos and many of them did not make any sense. Table 2 shows the number of tokens and types for raw word, surface, PAL lemma, MSA lemma, and gloss. The number of token instances is the same obviously for all categories in the table. The number of raw word unique types drops from 16,175 to 14,978 in the CODA compliant surface word. For example, the two raw words اتصلوا *AtSlwA* and اتصلو *AtSlw* have the same CODA surface word form اتصلوا *AtSlwA* ‘they called’. The additional drop in the number of unique lemma types is due to the lemma’s abstraction over a number of inflected forms, e.g., the surface words بيقولوا *byqwlwA* ‘they say’, قلنا *qlnA* ‘we said’ and قالها *qAlhA* ‘he said it’ allhave the same lemma: قال *qAl* ‘to say’.

Table 2 Number of tokens (instances) and types for different categories

Category	Tokens	Types
Raw word	55,960	16,175
Surface word	55,960	14,978
Lemma	55,960	7981
MSA lemma	55,960	8338
Gloss	55,960	10,189

Table 3 POS tag frequencies in Curras corpus

Tag	Frequency
Nouns (noun, noun_num, noun_quant)	19,574
Verbs (verb, verb_pseudo)	9386
Punctuations (punc)	9197
Adjectives (adj, adj_comp, adj_num)	3835
Prepositions (prep)	3299
Proper nouns (noun_prop)	3075
Pronouns (pron, pron_dem, pron_exclam, pron_interrog, pron_rel)	2641
Other particles (part, part_det, part_focus, part_fut, part_interrog, part_neg, part_restrict, part_verb, part_voc)	2146
Conjunctions (conj, conj_sub)	1325
Adverbs (adv, adv_interrog, adv_rel)	953
Others (interj, abbrev)	529

The Curras corpus is available for browsing online at: http://portal.sina.birzeit.edu/curras/

Table 3 shows the frequencies of POS tags in our corpus. We can observe that the most frequent tags in our corpus are the nominals (nouns, number nouns and quantifier nouns).

6 Palestinian CODA guidelines

This section extends the CODA E̦GY guidelines to include the PAL specifics. These extended guidelines are used to annotate each word in the corpus. As discussed earlier, one of the main challenges when building a corpus for a dialect is the lack of a conventional orthography or standard rules for writing in that dialect. The same DA word may be written in different forms, e.g., the word برضه *brDh* ‘also’ may be written as بردو *brdw*, برضو *brDw*, برضوا *brDwA*, or برده *brdh*. This creates a major problem in the annotation process as an annotator has to take into consideration all forms of a word, which is very difficult. To overcome this lack of standard orthography, a solution was proposed to define guidelines for the written Arabic dialects, called CODA (Habash et al. 2012a). Current guidelines of CODA are defined for Egyptian Arabic, and they are being extended to other dialects such as Tunisian Arabic (Zribi et al. 2014) and Algerian Arabic (Saadane and Habash 2015). Extending CODA to PAL requires observing and finding variations in PAL in terms of phonology, orthography, morphology, and lexicon. Next, we summarize these observations and propose a set of extensions to cover PAL in CODA. The full technical specification of our PAL CODA guidelines can be found in (Habash et al. 2016).

6.1 PAL observations and extensions to CODA guidelines

6.1.1 Etymological spelling of some root consonants

There are several sub-dialects within PAL that vary in terms of pronunciations of some root consonants (urban, rural, Bedouin, and Druze). For example, the MSA word قلب *qlb* 'heart' may receive four spellings that correspond to the four sub-dialectal pronunciations: قلب *qlb* /qalb/ , ألب *Âlb* /'alb/ , كلب *klb* /kalb/ , and جلب *jlb* /galb/. The original CODA guidelines state that: if a word's root is a cognate of an MSA root, then the root radicals are written using the corresponding MSA root radicals (this is only allowed for a specific set of consonant letters: ق *q,* ث *θ,* ذ *ð,*ض *D,* ظ *Ď,*ص *S* and ط *T*) (Habash et al. 2012a). As such all of the above variants must be spelled as قلب *qlb* to be CODA-compliant. In PAL CODA, we needed to add the consonant letter ك *k*, which is pronounced as /tš/ in some rural dialects. We also added the non-emphatic variants of the emphatic consonants (ض *D,* ظ *Ď,*ص *S* and ط *T*) since they are made emphatic in some sub-dialects. We provide in Table 4 a list of the possible variants for the extended PAL CODA consonant list, together with a number of positive and negative examples.

6.1.2 Levantine Arabic clitics

PAL speakers use clitics that are neither standard MSA nor included within the EGY CODA guidelines. We add the following PAL clitics to the CODA guidelines:

- The demonstrative proclitic هـ+ *h*+ 'this', e.g., هالبيت *b**h**Albyt* 'in this house'.
- The conjunction proclitic تـ+ *t*+ 'so as to', e.g., تيشوف ***t**yšwf* 'so that he can see'.

Additionally, we should note that some commonly attached clitics are required to be separated in EGY CODA, and we follow them in PAL also. These include the negation proclitic ما *mA* and the indirect object pronoun, which is written with the preposition ل+*l*+, e.g., the raw word متحكيلهاش *mtHkylhAš* 'don't tell her' should be written in CODA as the three words ما تحكي لهاش *mA tHky lhAš.*

6.1.3 Initial Hamza

Hamzas (glottal stops) at the beginning of the base word (i.e., the part of a word with no clitics attached) are written in different forms in dialectal content; some

Table 4 Variations of letters in PAL

CODA	Non-CODA variants	CODA examples	Non-CODA examples
ق *q*	ء ك *‘ k*	طريق *Tryq*	طريء *Try’*
		برتقان *brtqAn*	برتئان *brtŷAn*
		قال *qAl*	كال *kAl*
ك *k*	تش *tš*	كيف حالك *kyf HAlk*	تشيف حالتش *tšyf HAltš*
ث *θ*	س ت ط *s t T*	كثير *kθyr*	كتير *ktyr*
		ام كلثوم *Am klθwm*	ام كلسوم *Am klswm*
		ثور *θwr*	طور *Twr*
ذ *ð*	ظ ز د *Ď z d*	كذب *kðb*	كزب *kzb*
		ذل *ðl*	زل *zl*
ض د *D d*	ظ ز د *Ď z d*	ضابط *DAbT*	ظابط *ĎAbT*
ظ ز *Ď z*	ض ز ذ *D z ð*	ظل *Ďl*	ضل *Dl*
ص س *S s*	س ص *s S*	صاقع *SAqς*	ساقع *sAqς*
ط ت *T t*	ت ط *t T*	اللطف *AllTf*	اللتف *Alltf*
		فستان *fstAn*	فسطان *fsTAn*

people write it as in MSA, others drop it, and some write it incorrectly (by MSA standard). For example the proper name أحمد *ÂHmd* ‘Ahmad’ might be written as احمد *AHmd,* إحمد *ǍHmd,* or آحمد *ĀHmd*. The form of this Hamza is tied to a MSA spelling rule that distinguishes between Real Hamza and Hamzat-Wasl (همزة القطع وهمزة الوصل) based on the contextual phonology of words. This rule was used in EGY CODA as is. However, in our experience, it was hard to test and apply this rule in PAL as many phonological variants were possible. As such, we decided to relax this rule and allow Hamzas appearing at the beginning of the base word to be replaced with a bare Alif ‘ا’. Dropping the Hamza in this specific location will not, in most cases, create any ambiguity. Effectively, we consider the Hamza in the base word’s word-initial position as an optional diacritic variant (see Sect. 8.3).

6.1.4 Ta Marbuta

In PAL, there are different pronunciations of the Ta Marbuta (feminine singular) morpheme ة+ *+aħ*. For example, the word معلمة *mςlmħ* ‘teacher [fem.sing.]” can be written as معلمه *mςlmh* or معلمي *mςlmy.* Following the general CODA guidelines, we always write the Ta Marbuta as ة+ *+aħ*. Additional examples:

Non-CODA			CODA	
جميلي	jmyly	=>	جميلة	jmylħ
سياره	syArh	=>	سيارة	syArħ

Table 5 Part of the PAL exceptional list

CODA	Non-CODA variants	English
mAnyš مانيش	*mnyš* منيش	I'm not
AHnA احنا	*ǍHnA* إحنا	We
Anty انتي	*Ant Ǎnt* انت – إنت	You [2fs]
Antw انتو	*AntwA ǍntwA Ǎntw* انتوا -إنتوا – إنتو	You [2p]
Arbçħ اربعة	*Ârbçħ Arbçh Ârbçh* أربعة - اربعه - أربعه	Four
rH رح	*rAH* راح	Will [future particle]
hAðA هاذا	*hAĎA hAdA hAð hAD hADA* هاظا-هادا-هاذ-هاض-هاضا	This [ms]
jwA جوا	*jwħ jwh* جوة جوه	Inside
brDh برضو	*brDw brDwA brdh* برضو برضوا برده	Also

6.1.5 Epenthesis

All words ending with Consonant-Consonant (CC) clusters (that are not geminates) allow for a CiC epenthesized pronunciation, e.g., كلب /kalb/ and كَلِب/kalib/, ضَرب /Darb/ and ضَرِب/Darib/, and كَتَبْت/katabt/ and كَتَبِت /katabit/. Consonant clusters across words are broken up with an epenthetic vowel (الكسر لمنع التقاء السواكن). The vowel is typically /i/, e.g., /ibn/+/bla:d/ => /ibni bla:d/. In CODA we consider the non-epenthetic version as the base and only write it.

6.1.6 Lexical exceptions list

We created an extensive list of PAL words that have exceptional spelling, or whose CODA is not so obviously derivable given the rules. Table 5 includes some examples from the list of lexical exceptions.

6.2 Examples of PAL text in PAL CODA

Table 6 shows two examples of how PAL CODA guidelines are applied to PAL text.

7 Further annotation choices

While annotating our corpus using DIWAN (Al-Shargi and Rambow 2015), we came across scenarios where we were able to exploit the orthographic and morphological features shared between MSA, EGY, and PAL. In what follows, we describe these scenarios and discuss how our annotations were made.

No analysis words One scenario is when a word has no analysis in DIWAN in its raw word form; however, the annotator knows that this word has a similar word in MSA or EGY for which DIWAN can return an analysis. This eases and speeds up

Table 6 Examples of applying PAL CODA guidelines

Example 1	
Raw spelling	أنا حاسس انو بيوم عرسي أمي راح تحكيلي خد الزباله معك وانت طالع
	ÂnA HAss Anw bywm çrsy Âmy rAH tHkyly xd AlzbAlh mçk wAnt TAlç
PAL CODA	انا حاسس انه بيوم عرسي امي رح تحكي لي خذ الزبالة معك وانت طالع
	AnA HAss Anh bywm çrsy Amy rH tHky ly xð AlzbAlħ mçk wAnt TAlç
English	I feel that on my wedding day my mother will say to me “take the garbage with you when you go out.”
Example 2	
Raw spelling	وجراي وصارت ترمح وترمح ووكعت ببوجها عالدرلبست الفسطان وراحت جري عالكصر وشافها الأمير وحبها وركضو سو وفجـــــــــأة دكت الساعة طنعشج
	lbst AlfsTAn wrAHt jry çAlkSr wšAfhA AlÂmyr wHbhA wrkDw sw wfjÂħ dkt AlsAçħ Tnçš wjrAy wSArt trmH w trmH wwkçt bbwjhA çAldrj.
PAL CODA	لبست الفستان وراحت جري عالقصر وشافها الامير وحبها وركضوا سوا وفجأة دقت الساعة اطنعش وجراي وصارت ترمح وترمح ووقعت بابوجها عالدرج.
	lbst AlfstAn wrAHt jry çAlqSr wšAfhA AlAmyr wHbhA wrkDwA swA wfjÂħ dqt AlsAçħ ATnçš wjrAy wSArt trmH w trmH wwqçt bAbwjhA çAldrj.
English	She wore the dress and started running towards the castle then the prince saw her and fell in love with her, then they ran together and suddenly the clock pointed to twelve so she started running and running and she dropped her shoe on the stairs.

The transliteration is provided under the example the annotation for no-analysis words. The following are more specific cases under this scenario:

- *Word typos* In some cases, a word may contain typographical errors that could not have been intended by the writer. For example, the raw word فلطسيني *flTsyny* ‘Pale***ts***inian’ receives no analysis because it contains a typo. The annotator changes the spelling to فلسطيني *flsTyny* and is able to automatically fill all of the annotation metadata correctly.
- *Speech effects* It is not uncommon to find words spelled with elongated vowels that may signal surprise or excitement. For example, the raw word طوييييييل *Twyyyyyyl* ‘lo***ooooo***ng’ receives no analysis. The annotator changes the spelling to طويل *Twyl* and is able to automatically fill all of the annotation metadata correctly.
- *CODA non-compliant words* For example, the raw word المستشينة *Almstšynħ* ‘the poor one’ receives no analysis because it is not in a CODA compliant form. The annotator changes the spelling to المسكينة *Almskynħ* and is able to automatically fill all of the annotation metadata correctly.
- *Palestinian words* Words that do not exist in EGY or MSA return no analysis in DIWAN. Here, the annotator can annotate the word by filling each field in DIWAN; or alternatively, the annotator can look for a MSA or EGY word that has the same meaning and use it to automatically fill the metadata. The annotator will still need to validate and change some of the metadata as needed. For example, the PAL word سبيطار *sbyTAr* ‘clinic’ can be replaced by the MSA عيادة *ςyAdħ*, which is also a singular noun, although with different morphology. The gloss would be automatically filled correctly and only minor changes are needed.

Wrong analysis words Another interesting scenario is when DIWAN returns the wrong analysis for a word. The first step is for the annotator to determine if the correct analysis is available but not automatically selected. If the analysis is available, the annotator selects and fills the metadata automatically. However, in some cases, none of the available analyses are correct. Such cases require the annotator to carefully inspect the word to determine the best way to fix it. Specific cases include the following:

- *CODA non-compliant words* These are PAL words that are not CODA-compliant, but happen to match *other* MSA or EGY words. For example, the word كال *kAl* (CODA should be قال *qAl* ‘he said’), can be wrongly analyzed by DIWAN as the MSA verb for ‘measure’. The annotator changes the word to the CODA spelling and is able to fill the metadata automatically after selecting the desired analysis.
- *Palestinian words* PAL words that look like other words in EGY or MSA may be helpful in filling some of the metadata, e.g., the PAL verb ترمح *trmH* ‘she ran’ returns the MSA verb meaning ‘she pierced’. Only some of the metadata needed to be corrected in this case.

8 Corpus evaluation

This section presents three evaluations: a quantitative evaluation in terms of inter-annotator agreement and accuracy, a qualitative evaluation of differences in annotation produced by different annotators, and a detailed analysis of the CODA annotations.

8.1 Quantitative evaluation

8.1.1 Metrics

We performed a quantitative evaluation measuring the *accuracy* of our annotators, and their *inter-annotator agreement* (IAA). The *accuracy* of an annotator is defined as the degree of correctness of the annotations done by this annotator compared to a gold reference. We define a *gold reference* to be a set of annotations made and agreed by experienced annotators. The degree of the IAA is determined using the *Kappa coefficient, κ* (Cohen 1960; Di Eugenio and Glass 2004). This coefficient provides a good measure of agreement (Artstein and Poesio 2008) as it takes into consideration the agreement by chance along the observed agreement. The *Kappa coefficient κ* is widely regarded as a measurement of IAA, and has been used by several researchers to evaluate their work, such as the evaluation of the Irish Treebank (Lynn et al. 2012), evaluation of Hindi Treebank (Gupta et al. 2010), and evaluation of Basque corpus (Uria et al. 2009). In addition, it has been used as a measure of annotators' objectivity in different tasks; such as POS annotation (Mieskes and Strube 2006), discourse annotation in the GNOME corpus (Poesio 2004), and word sense disambiguation (Bruce and Wiebe 1998; Véronis 1998). The Kappa coefficient of measurement is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the observed agreement between annotators, and P_e is the expected agreement "agreement by chance" defined as the agreement between annotators obtained if they randomly assign tags while annotating. P_e is calculated as:

$$P_e = \sum_q \frac{n_{A_1 q}}{i} \times \frac{n_{A_2 q}}{i} = \frac{1}{i^2} \sum_q n_{A_1 q} \times n_{A_2 q} \tag{2}$$

where $n_{A_x q}$ is the number of words to which annotator A_x assigned tag q. *i* is the total number of annotated words.

8.1.2 Data

We selected three documents from our corpus, and asked two annotators (A1 and A2) to annotate them independently. The two annotators are the same people who annotated the whole Curras corpus. The three selected documents consist of 1529

Table 7 Inter-annotator agreement, disagreement, and Kappa values for different fields

Category	Agreement	Disagreement	Observed agreement	Kappa
Complex-prefix	1471	58	96	0.93
Stem	1380	149	90	0.88
Complex-Suffix	1394	135	91	0.83
Surface (CODA)	1437	92	94	–
Lemma	1390	139	91	–
$\text{Lemma}_{\text{MSA}}$	1268	261	83	–
Gloss	1345	184	88	–
POS	1389	140	91	0.88
Person	1484	47	97	0.92
Aspect	1506	23	98	0.96
Gender	1311	218	86	0.73
Number	1419	110	93	0.82

tokens (883 word types). After annotating the three documents, the two annotators were asked to meet, review and discuss their annotations, and to come up with verified and agreed annotations. These agreed annotations are then used as a *gold reference*. The gold reference and both independent annotations can be found in (Jarrar and Alrimawi 2015a). After that, we calculated the IAA by comparing both annotations against each other. The accuracy of each annotator was also determined by comparing it against the gold reference.

8.1.3 Results

Table 7 presents the inter-annotator agreement and Kappa values for different tags (e.g., POS, stem, prefix, etc.). The common interpretation of the Kappa value, as found in the literature (Landis and Koch 1977), segments the space between 0.0 and 1.0 into five equally distributed levels (0.2 width) corresponding to Slight, Fair, Moderate, Substantial and Almost Perfect. An absolute 0.0 value means there is no agreement. The categories complex-prefix and complex-suffix treat the multiple prefixes or suffixes with the Buckwalter POS Tag (BWT), respectively, as a single unit. The basic prefix/suffix categories are based on intersection over union of choices made by the two annotators.

Our results illustrate that the obtained Kappa coefficient for all categories is "almost perfect", except for Gender that has a "substantial" Kappa value. This reflects a high degree of inter-annotator agreement. The encountered disagreements will be discussed in the following subsection. The lexical items (Surface, Lemma, MSA Lemma and Gloss) are open classes, and as such we do not report Kappa for them. For further information regarding the calculations of Kappa values see this supplementary document (Jarrar and Alrimawi 2015b).

Table 8 shows the accuracy of each annotator, A1 and A2, which was obtained by comparing each of them against the agreed gold reference. This comparison

Table 8 Accuracy for annotators A1 and A2

Category	A1 accuracy	A2 accuracy
Complex-prefix	97.4	97.6
Stem	86.3	93.1
Complex-suffix	92.5	96.0
Surface (CODA)	90.8	97.8
Lemma	93.4	93.1
$\text{Lemma}_{\text{MSA}}$	87.5	87.2
Gloss	74.6	80.9
POS	94.1	95.7
Person	97.8	98.8
Aspect	99.3	99.1
Gender	86.2	96.9
Number	94.8	96.8

illustrates a high accuracy and a considerable degree of matching between both annotators and the gold reference. The low accuracy of the Gloss is due to simple disagreements and to the fact that this was not a focus for our annotators, specially that a gloss is an informal semantic denotation of a lemma. We report the numbers here for completeness; and we plan to investigate further in the future.

8.2 Qualitative evaluation

Although our quantitative analysis results illustrate high agreement between our annotators, there were some disagreements. We broadly classify the reasons of the most frequent mistakes and disagreements into the following groups:

- *Unintended annotator mistakes* In some cases the annotator mistakenly does not follow the annotation guidelines, or does not understand the meaning of the word in its context correctly. For example, the word ارجعلي *Arjçly* was mistakenly annotated as the imperfective verb meaning ‘I return to me’ by A1 while it should have been the command verb ‘return to me!’.
- *Using DIWAN tool* Although DIWAN provided great assistance by recommending annotations; however, this was also a source of errors, as the annotators sometimes forgot to correct the suggested annotations. For example, to annotate the PAL word عرض *çrD* ‘honor’, an annotator used the MSA شرف *šrf* but forgot to fully edit all the metadata fields. Additionally, incorrect annotations were made due to mistakes when writing in Buckwalter transliteration, such as writing *xls* instead of *xlS* ‘enough’.
- *Semantic ambiguity* In some instances, the two annotators had different understandings of a word in a certain context, and both understandings are possible.

- *Foreign words* In some infrequent cases, English words used in PAL (e.g., الامبالانس *AlAmbAlAns* ‘ambulance’) were tagged by A1 as a NOUN (correct), and A2 as FOREIGN (incorrect).
- *MSA lemmatization* Some words in PAL do not have exact equivalent words in MSA, but they might be illustrated by a closely related MSA word, a phrase, or a sentence to express their meaning. Such words cause disagreements between annotators, since the exact meaning in the context and the tag that should be used is not always clear. For example, the word اتاريه *Ataryh* ‘it seems that he is …’ has no direct equivalent word in MSA. Another example is the word تفو *tfw* (onomatopoeia for *spit*), which the annotators puzzled on how to give it an MSA lemma.

There were other minor disagreements related to missing one of the attributes (e.g., number or gender) when tagging a prefix or a suffix of a word, or having different diacritics for the lemma or MSA lemma. For example, one of the annotators tagged the suffix of the word فيكو *fykw* ‘in you [plural]’ as kw/PRON_2P while the other annotator tagged it as kw/PRON_2MP. Moreover, the lemma of the word ترغفة *trgfħ* ‘loaves of bread’ was written by A1 as “*riγiyf*” while A2 wrote it as “*raγiyf*”, where the only difference is in the diacritic as can be seen.

8.3 Analyzing CODA annotations

Recall that each word *w* in the corpus is re-written (i.e., transformed) into its surface *s,* according to the PAL CODA guidelines. This subsection provides an analysis on the use of these guidelines and the transformations made. Only 12.4 % of all words in the Curras corpus were changed when the annotators created their surface CODA form. This percentage suggests that the CODA guidelines are generally compatible with the choices people naturally make when they write, even without the existence of a standard orthography. The percentage is higher than the one reported for EGY by Eskander et al. (2013) for exact matching (24.5 %), but lower than their Alif/Ya normalized matching (10.3 %) (numbers reported here are on their blind test). This may be due to the simplification of Hamza spelling in PAL CODA. In this section, we present two sets of analyses of the kinds of spelling modifications our annotators did as part of creating the CODA annotations. The first analysis is an automatic analysis that classifies the character transformations; and the second is a more nuanced manual analysis that allows us to get additional insights in the CODA annotation task.

8.3.1 CODA character transformation statistics

To collect the CODA character transformation counts over the whole Curras corpus, we make use of the commonly used statistical word alignment tool GIZA++ (Och and Ney 2003); however, we apply it on the character level of pairs of raw and CODA words. The character level alignments allow us to identify insertions, deletions, and substitutions. See Fig. 3 for an illustrated example.

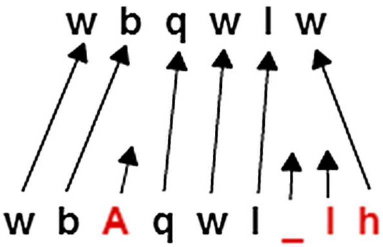


Fig. 3 Character-level alignment example. The underscore (_) indicates added space

The results presented in Table 9 show a coarse-grained analysis of the CODA modifications purely based on the observed character-level changes in the words. The percentages provided are of the set of CODA transformations. The most common CODA transformation has to do with the Hamza spelling. This is followed by the separation of various clitics. Ta Marbuta spelling follows with a solid 10 % plus. Transformations based on MSA root etymological spelling are also quite prominent if counted collectively (at least over 8 %).

8.3.2 CODA word transformation classification

To obtain a more fine-grained analysis of the CODA transformations, we selected a random sample of 600 CODA-transformed words and manually categorized them in three classes based on the reason for the transformation: (1) direct CODA guideline application (CODA) (2) typos in the text (TYPO) or (3) errors in the gold annotations (GOLD). Within each of these classes, we identified additional subclasses. The CODA guideline transformations were classified into (a) morphological form transformation (MORPH), (b) MSA root radical etymological spelling (ROOT), (c) pattern based spelling transformation (PATTERN), (d) splitting into to two words (SPLIT), and (e) acceptable variations of CODA covering mostly the Hamza at the beginning of the base word (CODA-VARIANT).

The TYPO transformations refer to CODA modifications that are not phonologically plausible as part of PAL. Within this scope, the typo corrections require deeper analysis of what might have caused the error in the first place in order to be able to classify it correctly. Typos that result from confusing a character with another that has a similar shape, like ح with خ (*H* with *x*), are classified as SHAPE. Moreover, typos that result from mistakenly deleting a character of the word are classified as DELETION. Typos that result from Hamza errors are classified as HAMZA. And finally typos that result from mistakenly typing a wrong character that is close to the correct character on the keyboard are classified as KEY, like typing "م" instead of "ك" (*m* instead of *k*), where the two keys are placed right next to each other in the Arabic keyboard. Since a number of Arabic keyboard characters share similar shapes and are also placed next to each other on the keyboard, KEY only refers to characters with different shapes.

GOLD errors are transformations that should not have been done. We specifically mark the subset of the GOLD transformations that are related to bad lexical transformation as LEX. Finally, in some very rare occurrences, the alignment tool

Table 9 Observed CODA character transformations

Modification phenomenon	Percentage	Examples
A ↔ Hamza	32.7	*ttAxr → ttÂxr, ÂHky → AHky, Ǎsmς → Asmς* تتاخر← تتأخر، إحكي←احكي، إسمع←اسمع
Splitting words	12.2	*qAltlh → qAlt_lh, mAntbhtš → mA_Antbhtš* قالتله ← قالت_له، مانتبهتش← ما_انتبهتش
h ↔ ħ	10.8	*klmh → klmħ, ςlyħ → ςlyh* كلمه ← كلمة، علية ← عليه
Inserting A	10	*bqlk → bAqlk* بقلك ← باقلك
Inserting y	7.2	*bHky → byHky* بحكي ← بيحكي
ý ↔ y	4.6	*tnAdý → tnAdy, mSTfy → mSTfý* تنادى ← تنادي، مصطفي ← مصطفى
w ↔ h	4.4	*šςrAtw → šςrAth* شعراتو ← شعراته
d → ð	2.4	*yAxdk → yAxðk* ياخدك ← ياخذك
k → q	2.3	*AlkSħ → AlqSħ* الكصة ← القصة
t → θ	1.8	*tlj → θlj* تلج ← ثلج
D ↔ Ď	1.5	*bnDl → bnĎl, HAĎr → HADr* بنضل ← بنظل، حاظر ← حاضر
A → ħ	0.9	*nmrA → nmrħ* نمرا ← نمرة
ŷ → q	0.43	*bnŷdr → bnqdr* بنئدر ← بنقدر
H → x	0.37	*bTyH → bTyx* بطيح ← بطيخ

Others	8.4	*tjςny → twjςny, lzyz → lðyð, yrDA → yrDý*
		تجعني ← توجعني، لزيز ← لذيذ، يرضا ← يرضى
		mςwš → mςhwš, ytfy → yTfy
		معوش ← معهوش، يتفي ← يطفي

we used (GIZA++) produced erroneous alignments; we classified these as ALIGN-ERROR.

Table 10 shows the results of the manual classification of the CODA transformations. CODA guideline-based modifications take up the majority of the transformations (86.54 %). CODA-MORPH and CODA-VARIANT are the two biggest classes. TYPO corrections are large but not unexpected. Among these errors, SHAPE cases are most common. The low rate of GOLD errors (2.8 %) is

Table 10 CODA transformation classifications

	Classification	Examples	Percentage	Total (%)
Transformations based on CODA Guidelines	CODA-MORPH	المدرسة ← المدرسه *Almdrsh* → *Almdrsħ*	36.67	86.54
	CODA-VARIANT	ارجع ← إرجع *Ǎrjς* → *Arjς*	25.2	
	CODA-ROOT	تقولي ← تكولي *tkwly* → *tqwly*	13.27	
	CODA-SPLIT	قلت_لي ← قلتلي *qltly* → *qlt_ly*	10.33	
	CODA-PATTERN	يادوب ← يدوب *ydwb* → *yAdwb*	0.67	
Typo Corrections	TYPO-SHAPE	الخارج ← الحارج *AlHArj* → *AlxArj*	5.67	10.33
	TYPO-DELETION	دايما ← ديما *dymA* → *dAymA*	2.83	
	TYPO-HAMZA	المسألة ← المسالة *AlmsAlħ* → *AlmsÂlħ*	1.5	
	TYPO-KEY	كبيرة ← طبيرة *Tbyrħ* → *kbyrħ*	0.33	
Gold Errors	GOLD-ELSE	وحصانة ← وحصانه *wHSAnh* → *wHSAnħ*	1.80	2.80
	GOLD-LEX	عرض ← شرف *ςrD* → *šrf*	1.00	
Other	ALIGN-ERROR	بالعزا ← بالفزآ *bAlfzĀ* → *bAlςzA* *aligning Ā with ς*	0.33	0.33

reassuring about the quality of the annotations and is consistent with the results presented in the previous two subsections.

9 Conclusion and future work

In this paper we presented Curras, the first annotated corpus for the Palestinian Arabic dialect. We described the various challenges we faced and solutions we took to produce balanced corpus of Palestinian Arabic, consisting of 55,960 tokens annotated with rich morphological and semantic information. While identifying text sources and managing the balance of corpus documents was a challenge, we were able to exploit many exiting resources and tools developed for MSA and other dialects, particularly Egyptian Arabic. We extended existing guidelines for Conventional Orthography of Dialectal Arabic to handle Palestinian Arabic. We also exploited tools for morphological analysis and disambiguation, such as MADIMARA. The use of these resource significantly helped us bootstrap our annotation effort. The challenge of finding good annotators and training them is not much different than any other annotation project in principles, but we believe that working on a primarily spoken dialect required more effort in training than if we were working on MSA annotations. Still, the evaluation of our annotators' performance shows a high degree of consistency and agreement. The Curras corpus is available for downloading and browsing online at: http://portal.sina.birzeit.edu/curras.

In the future, we plan to increase the size of our corpus to cover more of the various domains it contains, and target additional sub-dialects. We also plan to use this corpus to develop morphological analyzers and disambiguation systems for Levantine Arabic. Most importantly, we are currently working on linking the corpus with the Arabic Ontology (Jarrar 2006, 2011), which is a WordNet-like resource

(Miller et al. 1990) for Arabic, but with richer and more ontologically clean content. Each lemma in the corpus will be linked with its concept in the ontology. In this way the corpus will be enriched in different ways. Synonyms, glosses and other semantic features and relationships will also be automatically inherited from the ontology. This will enable us to additionally build an English-MSA-PAL lexicon.

Acknowledgments This work is part of our ongoing Curras project, funded by the Palestinian Ministry of Higher Education, Scientific Research Council. We wish to thank Owen Rambow, Ramy Eskander and Faisal Al-Shargi for their support with DIWAN and MADAMIRA. We would like to also thank Rami Asia for developing the Curras portal, Bahya Mustafa and Mohammad Dwaikat for their support during the annotation process, and Mahdi Arar for helpful conversations and fruitful discussions in the early stages of this work. Last but not least, we would like to thank the "Watan Aa Watar" actors for their support and for providing us with the scripts of their TV show.

References

Abdul-Mageed, M., & Diab, M. (2014). SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European language resources association (ELRA), Reykjavik, Iceland* (pp. 1162–1169).

Abdul-Mageed, M., Kübler, S., & Diab, M. (2012). Samar: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, *association for computational linguistics, Jeju, Korea* (pp. 19–28).

Alansary, S., Nagi, M., & Adly, N. (2007). Building an international corpus of Arabic (ICA): Progress of compilation stage. In *The 7th international conference on language engineering, Cairo, Egypt.*

Alkuhlani, S., & Habash, N. (2011). A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the association for computational linguistics: Human language technologies* (pp. 357–362).

Al-Sabbagh, R., & Girju, R. (2010). Mining the web for the induction of a dialectical Arabic lexicon. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10), European language resources association (ELRA), Malta* (pp. 288–293).

Al-Sabbagh, R., & Girju, R. (2012). YADAC: Yet another dialectal arabic corpus. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12), European language resources association (ELRA), Reykjavik, Iceland* (pp. 2882–2889).

Al-Shargi, F., & Rambow, O. (2015). DIWAN: A dialectal word annotation tool for Arabic. In *Proceedings of the second workshop on arabic natural language processing, association for computational linguistics, Beijing, China* (p. 49).

Al-Sughaiyer, I., & Al-Kharashi, I. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology, 55*(3), 189–213.

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics, 11*(2), 135–171.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596.

Attia, M. (2006). An ambiguity-controlled morphological analyzer for modern standard Arabic modelling finite state networks. In *Proceedings of the challenges of Arabic for NLP/MT conference, The British Computer Society, London, UK* (pp. 1–16).

Bakr, H. A., Shaalan, K., & Ziedan, I. (2008). A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *The 6th international conference on informatics and systems, (INFOS2008), Cairo University, Cairo, Egypt* (p. 72).

Beesley, K. R. (1996). Arabic finite-state morphological analysis and generation. In *Proceedings of the 16th conference on computational linguistics* (Vol. 1, pp. 89–94).

Bouamor, H., Habash, N., & Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European language resources association (ELRA), Reykjavik, Iceland* (pp. 1240–1245).

Bruce, R. F., & Wiebe, J. (1998). Word-sense distinguishability and inter-coder agreement. In *Proceedings of the empirical methods on natural language processing conference (EMNLP'98), association for computational linguistics, Granada, Spain 1998* (pp. 53–60).

Buckwalter, T. (2004). *Buckwalter Arabic morphological analyzer: Version 2.0*. LDC catalog number LDC2004L02. ISBN 1-58563-324-0.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Darwish, K. (2013). Arabizi detection and conversion to Arabic. arXiv preprint arXiv:1306.6755.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101.

Diab, M., Al-Badrashiny, M., Aminian, M., Attia, M., Dasigi, P., Elfardy, H., et al. (2014). Tharwa: A large scale dialectal Arabic-Standard Arabic-English lexicon. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, *European language resources association (ELRA), Reykjavik, Iceland* (pp. 3782–3789).

Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. In *LREC workshop on semitic language processing, Malta* (pp. 66–74).

Diab, M., Hacioglu, K., & Jurafsky, D. (2007). Automated methods for processing Arabic text: From tokenization to base phrase chunking. In: *Arabic computational morphology: Knowledge-based and empirical methods.* Kluwer/Springer.

Eskander, R., Al-Badrashiny, M., Habash, N., & Rambow, O. (2014). Foreign words and the automatic processing of Arabic social media text written in Roman script. *In Proceedings of the empirical methods on natural language processing conference (EMNLP'14), Doha, Qatar* (p. 1).

Eskander, R., Habash, N., Rambow, O., & Tomeh, N. (2013). Processing spontaneous orthography. In *Proceedings of the North American chapter of the association for computational linguistics: Human language technologies (NAACL HLT'13)*, Atlanta, Georgia (pp. 585–595).

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., et al. (1997). *CALLHOME Egyptian Arabic transcripts.* LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., & Buckwalter, T. (2009). Standard Arabic morphological analyzer (SAMA) version 3.1. In *Linguistic Data Consortium LDC2009E73.*

Gupta, M., Yadav, V., Husain, S., & Sharma, D. M. (2010). Partial parsing as a method to expedite dependency annotation of a Hindi treebank. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10), Malta* (pp. 1930–1935).

Habash, N. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies, 3*(1), 1–187.

Habash, N., Diab, M., & Rambow, O. (2012a). Conventional orthography for dialectal Arabic. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12), European language resources association (ELRA),* Istanbul, Turkey (pp. 711–718).

Habash, N., Jarrar, M., Alrimawi, F., Akra, D., Zalmout, N., Bartolotti, E., et al. (2016). *Palestinian Arabic conventional orthography guidelines*. Tech Report: Under preparation

Habash, N., Eskander, R., & Hawwari, A. A morphological analyzer for Egyptian Arabic. (2012b). In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology, association for computational linguistics, Montreal, Canada* (pp. 1–9).

Habash, N., & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd annual meeting on association for computational linguistics, Ann Arbor, Michigan, USA* (pp. 573–580).

Habash, N., & Rambow, O. (2006). MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics, association for computational linguistics, Sydney, Australia* (pp. 681–688).

Habash, N., Rambow, O., & Roth, R. (2009). MADA + TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt* (pp. 102–109).

Habash, N., & Roth, R. M. (2009). CATiB: The columbia Arabic treebank. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, association for computational linguistics, Beijing, China* (pp. 221–224).

Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal Arabic. In *proceedings of the North American chapter of the association for computational linguistics (NAACL'13)*, Atlanta, Georgia (pp. 426–432).

Habash, N., Soudi, A., & Buckwalter, T. (2007). On Arabic transliteration. In Arabic Computational (Ed.), *Morphology* (pp. 15–22). New York: Springer.

Herzallah, R. (1990). *Aspects of palestinian Arabic phonology: A nonlinear approach*. Ph.D., Cornell University, New York.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Washington, D.C.: Georgetown University Press.

Jarrar. (2006). Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *Proceedings of the 15th International World Wide Web Conference (WWW2006)*. Edinburgh, Scotland (pp. 497–503). ACM Press.

Jarrar. (2011). Building a formal Arabic ontology (Invited Paper). In *Proceedings of the experts meeting on Arabic ontologies and semantic networks. Alecso, Arab League*. Tunis, July 26–28, 2011.

Jarrar, M., & Alrimawi, F. (2015a). *Downloads.* http://sina.birzeit.edu/projects/curras/downloads. Accessed 18 Aug 2015.

Jarrar, M., & Alrimawi, F. (2015b). *Statistics and inter-annotator agreement calculations of the Palestinian dialect corpus—Curras*. www.jarrar.info/publications/JR15.pdf.

Jarrar, M., Habash, N., Akra, D., & Zalmout, N. (2014). Building a corpus for palestinian Arabic: A preliminary study. In *Arabic natural language processing* (*ANLP) workshop, at the conference on empirical methods in natural language processing (EMNLP 2014), Doha, Qatar* (p. 18).

Khalifa, S., Habash, N., Abdulrahim, D., & Hassan, S. (2016). A large scale corpus of Gulf Arabic. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'16). Portorož, Slovenia.*

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., & McLemore, C. (2002). Egyptian colloquial Arabic lexicon. In *LDC catalog number LDC99L22.*

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lynn, T., Cetinoglu, O., Foster, J., Ui Dhonnchadha, E., Dras, M., & van Genabith, J. (2012). Irish treebanking and parsing: A preliminary evaluation. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12), European language resources association (ELRA), Istanbul, Turkey* (pp. 1939–1946).

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., et al. (2006). Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*, *european language resources association (ELRA),* Genoa, Italy.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The Penn Arabic Treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools, Cairo, Egypt* (pp. 102–109).

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., & Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the 9th international conference on language resources and evaluation (LREC'14), Reykjavik, Iceland* (pp. 2348–2354).

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation.*

Mieskes, M., & Strube, M. (2006). Part-of-speech tagging of transcribed speech. In *Proceedings of the conference on language resources and evaluation (LREC'06)*, Genoa, Italy (pp. 935–938).

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–244.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Berlin: Springer Science & Business Media.

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2011). Arabic Gigaword fifth edition. In *LDC2011T11. Philadelphia: Linguistic Data Consortium.*

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., et al. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the conference on language resources and evaluation (LREC'14), Reykjavik, Iceland* (pp. 1094–1101).

Poesio, M. (2004). Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the workshop on discourse annotation, association for computational linguistics,* Barcelona, Spain.

Rafalovitch, A., & Dale, R. (2009). United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit* (Vol. 12, pp. 292–299).

Riesa, J., & Yarowsky, D. (2006). Minimally supervised morphological segmentation with applications to machine translation. In *Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA06)* (pp. 185–192).

Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In *Proceedings of the Arabic natural language processing (ANLP) workshop, Beijing, China* (p. 69).

Sajjad, H., Darwish, K., & Belinkov, Y. (2013). Translating dialectal Arabic to English. In *Proceedings of the association for computational linguistics, Sofia, Bulgaria*.

Salloum, W., & Habash, N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*.

Salloum, W., & Habash, N. (2013). Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In *proceedings of the North American chapter of the association for computational linguistics: Human language technologies (NAACL HLT'13)*, Atlanta, Georgia (pp. 348–358).

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

Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th conference of the association for machine translation in the Americas (AMTA)*, *Denver, Colorado.*

Shoufan, A., & Al-Ameri, S. (2015). Natural language processing for dialectical Arabic: A Survey. In *The Arabic natural language processing workshop 2015, Beijing, China.*

Smrž, O. (2007). Functional Arabic morphology. Formal system and implementation. *PhD Thesis, Charles University, Prague, Czech Republic.*

Smrž, O., & Hajic, J. (2006). The other Arabic treebank: Prague dependencies and functions. In *Arabic computational linguistics: Current implementations. CSLI Publications, 104*

Uria, L., Estarrona, A., Aldezabal, I., Aranzabe, M. J., De Ilarraza, A. D., & Iruskieta, M. (2009). Evaluation of the syntactic annotation in EPEC, the reference corpus for the processing of Basque. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 72–85). New York: Springer.

Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop. Herstmonceux Castle, UK* (pp. 2–4).

Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 1220–1229).

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., et al. (2012). Machine translation of Arabic dialects. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT'12).*

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., & Habash, N. (2014). A conventional orthography for Tunisian Arabic. *In Proceedings of the ninth international conference on language resources abd evaluation (LREC'14), European language resources association (ELRA), Reykjavik, Iceland* (pp. 2355–2361).