Maknuune: A Large Open Palestinian Arabic Lexicon

Shahd Dibas,† Christian Khairallah,‡ Nizar Habash‡
Omar Fayez Sadi,* Tariq Sairafy,* Karmel Sarabta,* Abrar Ardah*
†University of Oxford, ‡New York University Abu Dhabi
*University College of Educational Sciences - UNRWA

shahd.dibas@ling-phil.ox.ac.uk, christian.khairallah@nyu.edu, nizar.habash@nyu.edu

Abstract

We present Maknuune, a large open lexicon for the Palestinian Arabic dialect. Maknuune has over 36K entries from 17K lemmas, and 3.7K roots. All entries include diacritized Arabic orthography, phonological transcription and English glosses. Some entries are enriched with additional information such as broken plurals and templatic feminine forms, associated phrases and collocations, Standard Arabic glosses, and examples or notes on grammar, usage, or location of collected entry.

1 Introduction

Arabic is a collective of historically related variants that co-exist in a diglossic (Ferguson, 1959) relationship between a Standard variant and geographically specific dialectal variants. Standard Arabic (SA, العربية الفصحى) is typically used to refer to the older Classical Arabic (CA) used in Quranic texts and pre-Islamic poetry, all the way to Modern SA (MSA), the official language of news and culture in the Arab World. Dialectal Arabic (DA) is classified geographically into regions such as Egyptian, Levantine, Maghrebi, and Gulf. The dialects, which differ among themselves and SA, are the primary mode of spoken communication, although increasingly they are dominating in written form on social media. That said, DA has no official prescriptive grammars or orthographic standards, unlike the highly standardized and regulated MSA. In the realm of natural language processing (NLP), MSA has relatively more annotated and parallel resources than DA; although there are many notable efforts to fill gaps in all Arabic variants (Alyafeai et al., 2022).

In this paper, we focus on Palestinian Arabic (PAL), which is part of the South Levantine Arabic dialect subgroup. PAL consists of several sub-dialects in the region of Historic Palestine that vary in terms of their phonology and lexical choice (Jarrar et al., 2016). PAL, like all other DA, has been historically influenced by many languages, specifically, in its case, Syriac, Turkish, Persian, English and most recently Modern Hebrew (Halloun, 2019), as well as other Arabic dialects that came in interaction with PAL after the Nakba. While this research effort was originally motivated by the need to document and preserve the cultural heritage and unique identities of the various PAL sub-dialects, it has expanded to cover PAL’s ever-evolving nature as a living language, and provides a resource to support research and development in Arabic dialect NLP.

Concretely, we present Maknuune, a large open lexicon for PAL, with over 36K entries from 17K lemmas, and 3.7K roots. All entries include diacritized Arabic orthography and phonological transcription following Habash et al. (2018), as well as English glosses. Important inflectional variants are included for some lemmas, such as broken plural and templatic feminine. About 10% of the entries are phrases (multiword expressions) indexed by their primary lemmas. And about 67% of the entries include MSA glosses, examples, and/or notes on grammar, usage, or location of collected entry. To our knowledge, Maknuune is the largest open machine-readable dictionary for PAL. Maknuune is publicly viewable and downloadable.

We discuss some related work in Section 2, and highlight some PAL linguistic facts that motivated many of our design choices in Section 3. Section 4 presents our data collection process and annotation guidelines. We present statistics for our lexicon and evaluate its coverage in Section 5.

1/\maknūne/ is a PAL farming term that refers to an egg intentionally left behind in a specific location to encourage the chicken to lay more eggs in that location. We hope that the lexicon will encourage other researchers and citizen linguists to contribute to it.

2In this initial phase of Maknuune, we focus on the PAL sub-dialects spoken in the West Bank, an area with dialectal diversity across many dimensions such as lifestyle (urban, rural, bedouin), religion, gender, and social class.

3www.palestine-lexicon.org
2 Related Work

Linguistic Descriptions There are several linguistic references describing various aspects of PAL (Rice and Sa’id, 1979; Herzallah, 1990; Hopkins, 1995; Elihai, 2004; Talmon, 2004; Bassal, 2012; Cotter and Horesh, 2015). These are mostly targeting academics and language learners. We consulted many of these resources as part of developing our annotation guidelines.

Dialectal Corpora We can group DA corpora based on the degree of richness in their annotations. Some noteworthy examples of unannotated or lightly annotated corpora of relevance include the MADAR Corpus (Bouamor et al., 2018), comprising 2K parallel sentences spread across 25 dialects of Arabic, including PAL (Jerusalem variety) and the NADI corpus for nuanced dialect identification (Abdul-Mageed et al., 2021). The Shami Corpus (Abu Kwaik et al., 2018) includes 21K PAL sentences, and the Parallel Arabic Dialect Corpus (PADIC) contains 6.4K PAL sentences (Meftouh et al., 2015). In the spirit of genre diversification and wider coverage across dialects, El-Haj (2020) introduced the Habibi Corpus for song lyrics, which comprises songs from many Arab countries including all Levantine Arab countries.

Public and freely available morphologically annotated corpora are scarce for DA and often do not agree on annotation guidelines. A notable annotated dataset for PAL is the Curras corpus (Jarrar et al., 2016), a 56K-token morphologically annotated corpus. Other annotated Levantine dialect efforts include the Jordan Comprehensive Contemporary Arabic Corpus (JCCA) (Sawalha et al., 2019), the Jordanian and Syrian corpora by Alshargi et al. (2019), and the Baladi corpus of Lebanese Arabic (Al-Haff et al., 2022).

We consulted some of the public corpora as part of the development of Maknuune. However, most of the above datasets are based on web scrapes, which limits the amount of actual lemma coverage that they could attain.

Dialectal Lexicons Examples of machine-readable DA lexicons include the 36K-lemma lexicon used for the CALIMA EGY fully inflected morphological analyzer (Habash et al., 2012), based on the CALLHOME Egypt lexicon (Gadalla et al., 1997), and the 51K-lemma Egyptian Arabic Tharwa lexicon (Diab et al., 2014), which provides some morphological annotations.

The Palestinian Colloquial Arabic Vocabulary comprises 4.5K entries including expressions (Younis and Aldrich, 2021), and the MADAR Lexicon contains 2.7K entries dedicated to the Jerusalem variety of PAL, including lemmas, phonological transcriptions, and glosses in MSA, English and French (Bouamor et al., 2018).

In addition to the above there are a number of dictionaries for Levantine Arabic variants, e.g., Elihai (2004) (9K entries and 17K phrases for PAL), Halloun (2019) (for PAL), Freiha (1973) (ca. 5K entries for Lebanese Arabic), and Stowasser and Ani (2004) (15K entries for Syrian Arabic). These resources include base lemma forms, occasional plural forms, verb aspect inflections, and expressions; however, none of them are available in a machine-readable format, to the best of our knowledge.

The lexicon presented in this work strives to be a large-scale and open resource with rich entries covering phonology, morphology, and lexical expressions, and with a wide-ranging coverage of PAL sub-dialects. The lexicon may never be complete, but by making it open to sharing and contribution, we hope it will become central and useful to NLP researchers and developers, as well as to linguists working on Arabic and its dialects.

3 Linguistic Facts

In this section we present some general linguistic facts about PAL and highlight specific challenging phenomena that motivated many of our annotation decisions.

3.1 Phonology and Orthography

Like all other DA, and unlike MSA, PAL has no standard orthography rules (Jarrar et al., 2016; Habash et al., 2018). In practice, PAL is primarily written in Arabic script, and to a lesser extent in Arabizi style romanization (Darwish, 2014). Some of the variations in the written form reflect the words’ phonology, morphology, and/or etymological connections to MSA. Orthogonal and detrimental to the orthography challenge, PAL has a high degree of variability within it sub-dialects in phonological terms. We highlight some below, noting that some also exist in other DA.

Consonantal Variables A number of PAL consonants vary widely within sub-dialects. For example, the voiceless velar stop /k/ is affricated to the palatal /tʃ/ in many PAL rural varieties (Herzallah, 1990),

132
3.2 **Morphology**

Like other DA, PAL has a complex morphology employing templatic and concatenative morphemes, and including a rich set of morphological features: gender, number, person, state, aspect, in addition to numerous clitics. We highlight some specific morphological phenomena that we needed to handle.

**Ta Marbuta** The so-called feminine singular suffix morpheme, or Ta Marbuta (_aspect), is a morpheme that can be used to mark feminine singular nominals, but that also appears with masculine singular and plural nominals. Morphophonemically, it has a number of forms in PAL that vary contextually. First, in some PAL sub-dialects, the Ta Marbuta is pronounced as /a l a/ when preceded by an emphatic consonant, velars, and pharyngeal fricatives, e.g., مَكْلَةٌ baT~ah/lb a t t. /al/ ‘duck’; otherwise it realizes as /e l/ e.g., /صَبْهَةَ b i s~ih/lb i s s e l. In some northern PAL dialects, the /l e/ variant appears as /l i l/ in Idafa constructions, e.g., /لَيْبَةَ k a t b i l/. Finally, for some active participle deverbal nouns, the Ta Marbuta realizes as /a l a a l/ or /l i i l/ when followed by a pronominal object clitic, e.g., /كَتِبَةٍ k a h /

**Complex Plural Forms** Besides the common use of broken plural (templatic plural) in DA, we encountered cases of blocked plurals where a typical sound plural or templatic plural is not generated because another word form is used in its place (Aronoff, 1976). One example from Ramadin, is the plural form of the word جِبَالَةٌ /g a l b a/ ‘children’ [lit. weaklings]’.

3.3 **Syntax**

Previous research on Arabic dialects reveals that the syntactic differences between these dialects are considered to be minor compared to the morphological ones (Brustad, 2000). One particular challenging phenomenon we encountered is a class of nouns used in adjectival constructions, but violating noun-adjective agreement rules, which involve gender, number and rationality (Alkuhlani...
and Habash, 2011). For instance, the word *xiyxa* ‘removing the thorns from artichoke (Gundelia)’, and the phrase *baayt* ‘[lit. house]’ appears in compounds referring to celebrations, funerals, bathrooms, and whether or not a family has children (see the examples in Table 3).

**Collocations** As part of working on Maknuune, we encountered numerous collocations (words that tend to co-occur with certain words more often than they do with others). For example, the verbs used for trimming off the tough ends of some vegetables vary based on the vegetable: *ly Q a m m i 3 # b aa m y el* ‘trim off the tough ends of okra’, *ly Q a m m i 3 # b a a m y el* ‘trim off the tough ends of green beans’, *ly 3 a k k i b # 3 a k k u u b l* ‘remove the thorns from artichoke (Gundelia)’, and *ly t. a r l. i f # D u r a l* ‘cut the blossom ends of the maize stalks’.

**Compounds** We encountered many compositional and non-compositional compounds. Examples include *jawaAz sa far l J a w a a z # s a f a r l* ‘[lit. permission-of-travel, passport]’, which is also used in MSA. Some words appear in many compounds with a wide range of meaning, e.g., the word *baayt* ‘[lit. house]’ appears in compounds referring to celebrations, funerals, bathrooms, and whether or not a family has children (see the examples in Table 3).

**Synecdoches** It has been widely observed that PAL speakers use synecdoches in their dialects. Examples include the use of *lk oo m # l a 7 i m l* ‘[lit. a pile of meat]’, and *lk a b a a b ii shl* ‘[lit. plural of hair]’ to mean ‘children’.

**Euphemisms** PAL speakers use many euphemistic expressions. For example, in some villages in Nablus, the expression *ly oo m # t h a a n n a al* ‘[lit. the day he felt happy]’ to mean ‘the day he passed away’. In other areas in the West Bank, the phrase *3 e e n o # k a r i i m e l* ‘[lit. his eye is generous]’ to mean ‘one-eyed’; and the phrase *b e e t # h a a l a h i* ‘[lit. my aunt’s house]’ means ‘prison’.

### 4 Methodology

In this section, we discuss the methodology we adopted in data collection for Maknuune, as well as the guidelines we followed for creating the lexicon entries.

#### 4.1 Data Sources

The current work spans over five years of effort, and a large number of volunteering informants, linguistics students, and citizen linguists (over 130 people). The data was collected from many different sources.

First are interviews with (mostly but not entirely) elderly people who live in rural areas such as villages and towns or in refugee camps in the West Bank. The researchers went to the field and met with several people. They attended several social gatherings and participated in different events, e.g., weddings, funerals, field harvests, traditional cooking sessions, sewing, etc. They asked the language users several questions pertaining to the following themes: weddings, funerals, occupations, illnesses, cooking traditional dishes, plants, animals, myths, games, weather terms, tools and utensils, etc. They were particularly interested in documenting terms and expressions that are used mainly by the old generation.

Secondly, to achieve the needed balance in the lexicon, the researchers consulted an in-house balanced corpus, that contains ~40,000 words. The corpus comprises data that was transcribed from several recorded conversations that revolve around the same themes as above, written chats and texts, and some internet material (both written and spoken). Common words including verbs, adjectives,
adverbs, and function words (e.g., prepositions, conjunctions, particles) were taken from the balanced corpus. At a later stage in the development of Maknuune, we consulted with the Curras Corpus (Jarrar et al., 2016) to identify additional missing lemmas, with limited yield. We compare to Curras in terms of coverage in Section 5. All of the above was also supplemented by methodical rounds of well-formedness checking to improve consistency across all fields, i.e., diacritization, transcription, root validity, etc.

Finally, in addition to the previous two methods, the researchers employed their linguistic intuition skills, knowledge of Palestinian Arabic (as native speakers) and the knowledge of the language users to provide additional word classes and multiword expressions that are associated with the existing lemmas.

It should be noted that whether an MSA lemma cognate of a PAL lemma (with similar or exact pronunciation, or meaning) existed was not considered a factor in including the PAL lemma in the lexicon. We focused on creating a representative sample of PAL including all its sub-dialects.

4.2 Lexical Entries

Each entry in the Maknuune lexicon consists of six required and three optional fields. The six required fields are the Root, Lemma, Form, Transcription, POS & Features, and English Gloss. The optional fields are the MSA Gloss, Example and Notes. Figure 1 presents an example of a number of entries coming from the same root.

4.2.1 Root, Lemma, and Form

The Root, Lemma and Form represent three degrees of morphological abstraction. The root in Arabic in general is a templatic morpheme that interdigitates with a pattern or template to form a word stem that can then be inflected further. Roots are very abstract representations that broadly define the morphological family a word belongs to at the derivational and inflectional level. Lemmas on the other hand are abstractions of the inflectional space that is limited by variations in the morphological features of person, gender, number, aspect, etc. Lemmas are the central entries of the lexicon. Forms are base words (i.e., without clitics) that are inflected in a specific way. We follow the same general guidelines of determining lemmas as used in large Arabic morphological analyzers (Graff et al., 2009; Habash et al., 2012; Khalifa et al., 2017). There are of course some constructions that have grammaticalized into new lemmas, e.g., ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَан ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْنَان ُعُمْn
symbols we introduced. All of these symbols are to be presented in upper case, while normal CAPHI symbols are in lower case. The new CAPHI++ symbols represent specific sets of mostly two variants in common use in different PAL sub-dialects. For example, instead of including four entries for the word /kālam/ (/lq a l a m/, /lk a l a m/, /l2 a l a m/, /lg a l a m/), we only provide one form (/Q a l a m/). Exceptional usages that do not conform to the specific generalizations of the CAPHI++ cover symbols are listed independently, e.g., a second entry for the above example is provided for the Beit Fajar pronunciation of /tsh a l a m/.

We acknowledge that the transcriptions provided may not represent the full breadth of PAL sub-dialects. We make our resource open so that additional forms and variants can be added in the future, as needed.

### 4.2.3 POS and Features

The analysis cell in every entry indicates the POS and features of the word form. We use 35 POS tags based on a combination of previously used POS tags in Arabic NLP (Graff et al., 2009; Pasha et al., 2014; Khalifa et al., 2018). Our closest relative is the tagset used by (Khalifa et al., 2018) for work on Emirati Arabic annotation. See the full list of POS tags in Table 6 in Appendix A. However, we extend their POS list with three tags: ADJ/NOUS (for adjectives with exceptional agreement), NOUN_ACT (active participle deverbal noun), and NOUN_PASS (passive participle deverbal noun).

For features, we use MS (masculine singular), FS (feminine singular), and P (plural) for nominals, and P (perfective), I (imperfective) and C (command) for third masculine singular verb forms only.

### 4.2.4 Phrases

In addition to basic word forms, we overload the use of the form cells to list phrases (multword expressions, collocations, and figures of speech) that are paired with the lemma. In such cases, the POS:Features cell is given the POS of the lemma, with the extension PHRASE, e.g., line (d) in Table 1, and Table 3.

### 4.2.5 Gloses, Examples and Notes

We provided the English gloss equivalents of all the PAL words. The MSA gloss was provided for about a third of the entries at the time of writing. In cases where no single word in MSA or English can encode a culturally specific concept, the annotators translated the whole situation/concept. For example, in Ramadin, there are two words for

---

Table 2: The CAPHI++ symbols set and its expanded CAPHI symbols, with examples.

| CAPHI++  | CAPHI   | CAPHI Transcription | CODA | CAPHI++ Transcription |
|----------|---------|---------------------|------|------------------------|
| Q        | kq2g    | k qalam / qalam /   |      |                         |
| D        | dh      | dhab / dh / dh b   |      |                         |
| J        | j j     | r j / j j j a l     |      |                         |
| Z        | z dh    | zan b / dh anz     |      |                         |
| T        | t th    | tim m / th m       |      |                         |
| S        | st      | haw ra / saw ra    |      |                         |
| Z.       | z dh    | 2 a z un / 2 a h    |      |                         |
| D.       | d dh    | b ec e / b dh      |      |                         |
| K        | k tsh   | ke e f / tsh see f  |      |                         |

Table 1: Eight entries from Maknuune that share the same root, and are paired with four distinct lemmas.
Entries 17,369 36,302 32,759 3,543

Table 3: Examples of NC compounds in Maknuune for the lemma بيت ‘house’.

‘baby camel’ depending on its age: دلولُ ذَلَّطٍ الدُّلْلُ ُمْلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَلَل*al-Haff et al. (2022) describe a revised version of that corpus, but it was not made available at the time of writing.

Table 4: POS type and entry statistics in Maknuune.

5 Coverage Evaluation

We approximate the coverage of our lexicon by comparing it with the Curraas corpus (Jarrar et al., 2016), the largest resource available for PAL. Since Curraas is a corpus and our resource is a lexicon, the analysis is carried out in such a way to account for that difference. We present next some high-level corpus statistics and then a detailed comparison between Maknuune and Curraas. Then, we provide some comparison between Maknuune and the lexicons of two morphological analyzers for MSA and EGY.

5.1 Maknuune & Curraas Statistics

Maknuune POS Types Table 4 shows some basic statistics about Maknuune, dividing entries across four basic POS types (see Table 6). Maknuune has about three times more verb entries than verb lemmas, reflecting the fact that almost each verb appears in all three aspects (perfective, imperfective, and command) in third person masculine singular form. Similarly for nominals (nouns, adjectives, etc.), the ratio of 1.2 forms per lemma reflects the inclusion of plural entries for many
The Curras Lexicon In order to compare Maknuune with Curras, we extract a lexicon, henceforth Curras Lexicon, out of the Curras corpus by unifying its entries based on lemma, inflected form, POS, and grammatical features (for Curras, aspect, person, gender, and number). We compare the Curras Lexicon to Maknuune in Table 5.

Firstly, Curras does not include roots; and although it is a corpus, it does not identify phrases in the way Maknuune does. As such, we do not compare them in those terms in Table 5.

Secondly, by virtue of being a lexicon, Maknuune possesses more unique lemmas, weighing in at 17,369 lemmas taking POS into account (lemma:POS), while the total number of inflected forms is at 32,759, both of which are about 50% more than in the Curras Lexicon. This clearly showcases Maknuune’s richness in terms that go beyond the day-to-day language that one sees frequently in corpora like Curras. In contrast, Curras being a corpus, its extracted lexicon showcases a greater inflectional coverage with 224 unique word analyses as opposed to 76 for Maknuune.

Finally, as inferable from the difference between the number of unique lemmas and lemma:POS, 548 lemmas are associated to more than one POS in Maknuune.

5.2 Corpus Coverage Analysis

In the interest of estimating how well our lexicon would fare with real-world data, we perform an analysis between the Curras and Maknuune lemmas, to see how many of the Curras lemmas Maknuune actually covers. From an initial investigation, we note that there are numerous minor differences that need to be normalized to ensure a more meaningful evaluation. As such, we first pre-process all lemmas (in both lexicons) by stripping the sukun diacritic, stripping all the فتحة (i) and دل (a) diacritics if they appear before the آ ح م رة وصل أ, converting the Ä to Ä, and stripping the كسرة (i) and دل (a) diacritics if they appear before ㅎ. We then compare all the annotated lemma:POSType in Curras (56,004 tokens and 8,315 normalized types) to the lemmas in Maknuune.

We exclude 12,673 (23%) of the tokens pertaining to punctuation, digits and proper noun POS, none of which were especially targeted by Maknuune. Of the remaining 43,331 entries, 49% have exact match in Maknuune. We sample 10% of the unique entries with no exact match (433 types and 1,965 tokens), and manually annotate them for their mismatch class. We found that 74% of all the sampled types (80% in tokens) are actually present in Maknuune, but with slight differences in orthography mainly in the presence or absence of diacritics but also some spelling conventions. For about 20% of sampled types (17% in tokens), the lemma type is not one that we targeted such as foreign words and proper nouns that are differently labeled in Curras, or MSA words. Finally, 6% of sampled types (3% in tokens) are entries that are admittedly missing in Maknuune and can be added.

This suggests that we have very good coverage although the annotation errors and differences make it less obvious to see. A simple projected estimate assuming that our 10% sample is representative would suggest that Maknuune’s coverage of Curras’ lexical terms (other than proper nouns and punctuation) is close to 94% (97% in token space); however a full detailed classification would be needed to confirm this projection.

5.3 Overlap with MSA and EGY

In this section we conduct an evaluation similar to the one carried out in Section 5.2 but with an MSA lexicon (Calima<sub>MSA</sub>), and an Egyptian Arabiclex-
icon (Calima_{EGY}). The analysis reveals that 44% of Maknuune overlaps with Calima_{MSA} at the lemma:POSType level (63% if all entries are dediacritized), and that 49% of Maknuune overlaps similarly with Calima_{EGY} (75% dediacritized). Taking into account that Maknuune spelling follows the CODA* guidelines, the analysis suggests that the 37% of Maknuune lemma:POSTypes, which do not exist in the MSA lexicon we used, are heavily dialectal. The overlap with EGY is predictably higher, and the 25% of Maknuune lemma:POSTypes (dediacritized) not existing in EGY highlights the differences between the two dialects despite their many similarities.

5.4 Observations on Lexical Richness and Diversity

The quantitative analyses we presented above allow us to see the big picture in terms of lexical richness and diversity in Maknuune and its complementarity to existing resources. However, we acknowledge that such an approach misses a lot of details that are collapsed or lost when ignoring subtle differences in semantics, phonology and morphology.

We first point at homonyms showing semantic changes and spread, such as /a'awi/ 'small olives that people find hard to pick’ in some villages in Palestine and ‘ducks’ in both MSA and PAL, /'b a t. t./ which means ‘very small olives that people find hard to pick’ in some villages in Palestine and ‘ducks’ in both MSA and PAL, and /2 aa kh r e l/ which means ‘desserts’ in Nablus and ‘the Day of the Judgment’ in both MSA and PAL, albeit with a different pronunciation. Clearly, additional entries are needed to mark these differences.

Furthermore, the majority of the entries in Maknuune are actually pronounced differently from MSA even if spelled the same without diacritics and thus warrant entries of their own, with clear phonological specifications.

Finally, if we consider morphology (which is not modeled here per se), many PAL lemmas that have MSA lemma cognates are actually inflected differently, e.g., /mad/ ‘extend; stretch’ (in PAL and MSA), has different inflections for some parts of the paradigm: the 2nd person masculine plural is /mad~aytuwa/ in PAL and /madad.tum/ in MSA. Hence, each lemma in our lexicon heads a morphological paradigm which differs from its MSA counterpart.

6 Conclusion and Future Work

We presented Maknuune, a large open lexicon for the Palestinian Arabic dialect. Maknuune has over 36K entries from 17K lemmas, and 3.7K roots. All entries include Arabic diacritized orthography, phonological transcription and English glosses. Some entries are enriched with additional information such as broken plural and templatic feminine forms, associated phrases and collocations, Standard Arabic glosses, and examples or notes on grammar, usage, or location of collected entry.

In the future, we plan to continue to expand Maknuune to cover more PAL sub-dialects, more entries, and richer annotations, in particular for locations of usage, and morpholexical features such as rationality. We hope that by making it public, more researchers and citizen linguists will help enrich it and correct anything missing in it.

We also plan to make use of Maknuune as part of the development of larger resources and tools for Arabic NLP. The phonological transcriptions can be helpful for work in speech recognition and the morphological information for developing morphological analyzers and POS taggers. Furthermore, we plan to utilize Maknuune to develop pedagogical applications to help teach PAL to non-Arabic speakers and to children of Palestinians in the diaspora.

Acknowledgments

We would like to thank Prof. Jihad Hamdan, Muhammed Abu Odeh, Adnan Abu Shamma, Issra Ghazzawi and Kazem Abu-Khalaf for the helpful discussions.

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdeRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

\footnote{For MSA, we compared with the calima-msa-s31_0.4.2.utf8.db version (Taji et al., 2018) based on SAMA (Graff et al., 2009) and for EGY we only compared to the calima-egy-c044_0.2.0.utf8.db based on Habash et al. (2012). For EGY, only CALIMA analyses entries are selected.}

\footnote{The shadda (∼) is not included in dediacritization.
Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the Language Resources and Evaluation Conference, pages 769–778, Marseille, France. European Language Resources Association.

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

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, irani and Moroccan. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 137–147, Florence, Italy. Association for Computational Linguistics.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2022. Masader: Metadata sourcing for Arabic text and speech data resources. In Proceedings of the Language Resources and Evaluation Conference, Marseille, France.

Mark Aronoff. 1976. Word formation in generative grammar. Linguistic Inquiry, Monograph one, The MIT press.

Ibrahim Bassal. 2012. Hebrew and Aramaic Substrata in Spoken Palestinian Arabic. Mediterranean Language Review, 19:85–104.

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

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

William Cotter and Uri Horesh. 2015. Sociolinguistics of Palestinian Arabic. Encyclopedia of Arabic Language & Linguistics.

Kareem Darwish. 2014. Arabizi Detection and Conversion to Arabic. In Proceedings of the Workshop for Arabic Natural Language Processing (WANLP), pages 217–224, Doha, Qatar.

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

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1318–1326, Marseille, France. European Language Resources Association.

Yohanan Elilhai. 2004. The olive tree dictionary: A transliterated dictionary of conversational Eastern Arabic (Palestinian). Minerva Jerusalem.

Charles F Ferguson. 1959. Diglossia. Word, 15(2):325–340.

Anis Freiha. 1973. Dictionary of Non-Classical Vocables in the Spoken Arabic of Lebanon. Librairie du Liban.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

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

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In Proceedings of the Workshop of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON), pages 1–9, Montréal, Canada.

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

Moin Halloun. 2019. An etymological lexicon of foreign words in Palestinian Arabic : Arabic-Arabic-English : the influence of Greek, Pahlavi, Latin, Persian Syriac, Ottoman language and modern languages in the Palestinian dialect. Bethlehem: Bethlehem University, The Institute of Oral Cultural Heritage of the Palestinians.

Rukayyah S Herzallah. 1990. Aspects of Palestinian Arabic phonology: A nonlinear approach. Cornell University.
A  POS Type Mapping and Examples

| POS Type | Examples |
|----------|----------|
| Nominal  |         |
| ADJ      | أسد، سود، سودن          |
| ADJ_COMP | أثيرو، أثيرو، أثيرو          |
| ADJ/NOUN | فإله، إله، إله          |
| NOUN     | وَلَدَةٌ، وَلَدَةٌ، وَلَدَةٌ          |
| NOUN_ACT | تَفْتَرِةٌ، تَفْتَرِةٌ، تَفْتَرِةٌ          |
| NOUN_PASS | كَمْ يَضْمَكَ كَمْ يَضْمَكَ كَمْ يَضْمَكَ          |
| NOUN_QUANT | كِلُّي، كِلُّي، كِلُّي          |
| VERB     | البنِّي، الهَلَكُ          |
| VEB_PSEUDO |         |
| PROPER   |         |
| ABBREV   | اوَلَ، كَبِّيَةٌ، كَبِّيَةٌ          |
| ADJ_NUM  | هوَلُ، هَلَا، هَلَا          |
| ADV      | ماذا رَجَعَ لَبَغِيَةٌ          |
| ADV_INTEGROG | كَمْ فَهَلَّ فَهَلَّ          |
| ADJ_REL  | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| CONJ     | لا، لا، لا          |
| CONJ_SUB | كَمْ صَبَّ غُلَاثَ كَمْ صَبَّ غُلَاثَ          |
| INTERJ   | لا، لا، لا          |
| NOUN_NUM | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART     | وَلَدَةٌ، وَلَدَةٌ، وَلَدَةٌ          |
| PART_DET | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_FOCUS | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_FUT | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_INTERROG | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_NEG | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_PROG | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_RESTRICT | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PART_VOC  | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRP      | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRON     | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRON_EXCLAM | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRON_DEM | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRON_INTEGROG | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| PRON_REL  | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| VERB_NUM | كَمْ أَثَّرَ كَمْ أَثَّرَ          |
| VERB_PSEUDO | كَمْ أَثَّرَ كَمْ أَثَّرَ          |

Table 6: Mapping of part-of-speech (POS) types to POS tags used to annotate base words in Maknuune, and associated examples.