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 مكنونة, 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** مكنونة,[1] a large open lexicon for PAL, with over 36K entries from 17K lemmas, and 3.7K roots.[2] 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.[3]

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] مكنونة /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.

[2] In 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.

[3] www.palestine-lexicon.org

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 131 - 141

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 /*tsh*/ in many PAL rural varieties (Herzallah, 1990),

e.g., كَيف *kayf* 'how' appears as */k ee f/* (urban) or */tsh ee f/* (rural).[4] Similarly, the MSA voiceless uvular stop */q/* in the word قَلْب *qal.b* 'heart' is realized either as glottal stop */2 a l b/* in urban dialects, as a voiceless velar stop */k a l b/* in rural dialects, or a voiced velar stop */g a l b/* in Bedouin dialects (Herzallah, 1990). It should be noted that there are some exceptions that do not conform to the above generalizations. For example, in Beit Fajjar,[5] the word قَهْوَة *qah.waħ* 'coffee' typically varying elsewhere as */{2,q,g,k} a h w e/* is realized as */tsh h ee w a/*. Moreover, some words do not have varying pronunciations such as عْقَال ʕ.*qaAl* */3 g aa l/* 'Egal headband'.

Monophthongization Some PAL diphthongs shift to different monophthongs in different locations. For example the */a y/* diphthong in شَيخ *šayx* */sh a y kh/* 'Sheikh' shifts often to */ee/* (*/sh ee kh/*), but also to */ii/* (*/sh ii kh/*).[6] Following the CODA* guidelines for diacritizing DA (Habash et al., 2018), we spell the */oo/* and */ee/* sounds using وَ *aw* and يَ *ay* (without a *sukun* on the و *w* or ي *y*), respectively, e.g., كَوم *kawm /k oo m/* 'pile' and بَيت *bayt /b ee t/* 'house'.

Metathesis In some rural dialects in villages near Tulkarem, Jenin and Ramallah, there are words with consonant pairs within a syllable that appear in a different order than is the norm in PAL, e.g., a word like كَهْرَبَا *kah.rabaA /k a h r a b a/* 'electricity' realizes as */k a r h a b a/*.

Epenthesis PAL exhibits systematic epenthesis of the */i/* or */u/* sounds producing paired word alternations such as */b a 3 d/* and */b a 3 i d/* for بعد 'still;after' or */kh u b z/* and */kh u b u z/* or */kh u b i z/* (in different sub-dialects) for خبز 'bread'. We opted to use the fully epenthesized forms in the lexicon, i.e., بَعِد *baʕid*, خُبُز *xubuz*, and خُبِز *xubiz*, for the above mentioned examples.

[4]Arabic orthographic transliteration is presented in the HSB Scheme (italics) (Habash et al., 2007). Arabic script orthography is presented in the CODA* scheme, and Arabic phonology is presented in the CAPHI scheme (between /../) (Habash et al., 2018).

[5]A Palestinian town located 8 kilometers south of Bethlehem in the West Bank.

[6]In the Palestinian village of Ramadin, near Hebron in the West Bank.

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 (ة ħ), 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/* when preceded by an emphatic consonant, velars, and pharyngeal fricatives, e.g., بَطَّة *baT∼aħ /b a t. t. a/* 'duck'; otherwise it realizes as */e/*, e.g., بِسّة *bis∼iħ /b i s s e/*. In some northern PAL dialects, the */e/* variant appears as */i/*; and in some southern PAL dialects, the distinction is gone and all Ta Marbutas are pronounced */a/*. Second, the Ta Marbuta turns into its allomorph */i t/* in *Idafa* constructions, e.g., */b i s s i t/* 'the/a cat of'. Finally, for some active participle deverbal nouns, the Ta Marbuta realizes as */aa/* or */ii t/* when followed by a pronominal object clitic, e.g., كَاتْبَاه *kaAt.baAh /k aa t b aa (h)/* or كَاتْبِيْتُه *kaAt.biy.tuh* or */k a t b ii t u (h)/* 'she wrote it'.

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 عَيِّل ʕ*ay∼il /3 a y y i l/* 'child [lit. dependent]', which is blocked by the word form ضْعُوف *D.ʕuwf /dh. 3 uu f/* '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 خِيخَة *xiyxaħ /kh ii kh a/* 'weak/lame' does not typically agree with the nouns it modifies unlike a normal adjective such كْبِير *k.biyr /k b ii r/* 'old [human]/large [nonhuman]'. So, the words سِيَّارَة *siy∼aAraħ* 'car [f.s.]', عُرُس *ςurus* 'wedding [m.s.]', and نَاس *naAs* 'people [m.p]' can all be modified by خِيخَة *xiyxaħ*; however, they need three different forms of كْبِير *k.biyr*: كْبِيرِة *k.biyriħ*, كْبِير *k.biyr*, and كْبَار *k.baAr*, respectively. We mark the POS of such nominals as ADJ/NOUN in our lexicon, as it is a class that deserves further study.

3.4 Figures of Speech and Multiword Expressions

PAL has a rich culture of figures of speech and multiword expressions (compounds, collocations, etc.) that has not been well documented. We highlight some phenomena that we cover in Maknuune.

Collocations As part of working on Maknuune, we encountered numeorus 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: يْقَمِّع بَامْيِا */y Q a m m i 3 # b aa m y e/* 'trim off the tough ends of okra', يْقَرِّم فَاصُولْيَا */y q a r r i m # f aa s. uu l y a/* 'trim off the tough ends of green beans', يْعَكِّب عَكُّوب */y 3 a k k i b # 3 a k k uu b/* 'remove the thorns from artichoke (Gundelia)', and يْطَرْطِف ذُرَة */y t. a r t. i f # D u r a/* 'cut the blossom ends of the maize stalks'.

Compounds We encountered many compositional and non-compositional compounds. Examples include جَوَاز سَفَر *jawaAz safar /J a w aa z # s a f a r/* '[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 بَيت *bayt* '[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[7] in their dialects (Seto, 1999). Examples include the use of كَوم لَحِم */k oo m # l a 7 i m/* '[lit. a pile of meat]', and كَبَابِيش */k a b aa b ii sh/* '[lit. plural of hair]' to mean 'children'.

Euphemisms PAL speakers use many euphemistic expressions. For example, in some villages in Nablus, the expression يَوم تْهَنَّى */y oo m # t h a n n a/* '[lit. the day he felt happy]' to mean 'the day he passed away'. In other areas in the West Bank, the phrase عَينُه كَرِيمِة */3 ee n o # k a r ii m e/* '[lit. his eye is generous]' to mean 'one-eyed'; and the phrase بَيت خَالْتِي */b ee t # kh aa l t 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,

[7] A figure of speech in which a term for a part of something is used to refer to the whole, or vice versa.

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) exists 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., عَشَان *ςašaAn* can be treated as the noun شَان *šaAn* 'situation;status' with a proclitic, or the subordinating conjunction meaning 'because'.

For nouns and adjectives, we provide the lemma in the masculine singular form, unless it is a feminine form that does not vary in gender, in which case it is provided in the feminine singular. Very infrequently, some nouns only appear in plural form, which become their lemma, e.g. أَوَاعِي *ÂawaAçiy* */2 a w aa 3 i/* 'clothes'. We do not list the sound plural and sound feminine inflections of nouns and adjectives. However, broken plurals and templatic feminine forms are provided and linked through the same lemma as the singular form.

For verbs, we provide the lemmas in the third masculine singular perfective form as is normally done in Arabic lexicography. We provide three forms linked to the lemma: the third masculine singular perfective, the third masculine singular imperfective, and the second person masculine imperative (command) forms. These are provided for completeness to identify the basic verbal inflectional paradigm (albeit, not completely).

These three representations are provided in Arabic script. Since PAL does not have an official standard orthography, we intentionally decided to follow the Conventional Orthography for Dialectal Arabic (CODA*) (Habash et al., 2018). In addition to being used in developing Curras (Jarrar et al., 2016), CODA* has been adopted by a website for teaching PAL to non-native speakers.[8]

4.2.2 Transcription with CAPHI++

One of CODA*'s limitations is that it abstracts over some of the phonological variations. As such, we follow the suggestions by Habash et al. (2018) to use a phonological representation, CAPHI, to indicate the specific phonology of the entries. CAPHI, which stands for Camel Phonetic Inventory is inspired by the International Phonetic Alphabet (IPA) and Arpabet (Shoup, 1980), and is designed to only use characters directly accessible on the common keyboard to ease the job of annotators.

Owing to the phonological variations that are found in PAL, we extended CAPHI's symbol set with *cover phonemes* that represent a number of possible interchangeable phones. We call our extended set CAPHI++. Table 2 presents the new 9

[8] https://www.palestinianarabic.com/

	Root	Lemma	Form	Transcription	POS:Features	English	MSA	Example	Notes
(a)	ت.ف.ح	تُفّاح	تُفّاح	t u f f aa 7	NOUN:MS	apples	تُفّاح	كِلو تُفّاح أقل شي رح يكلفك 8 شيقل	Collective Noun
(b)	ت.ف.ح	تُفّاحَة	تُفّاحَة	t u f f aa 7 a	NOUN:FS	apple	تُفّاحَة	كان الصحن قدامي فتناوَلِت تُفّاحَة بس طلعت مدودة	Unit Noun
(c)	ت.ف.ح	تُفّاحَة	تَفافِيح	t a f a f ii 7	NOUN:P	apple			
(d)	ت.ف.ح	تُفّاحَة	تُفّاحِة آدَم	t u f f aa 7 i t # 2 aa d a m	NOUN:PHRASE	Adam's apple		شايف تُفّاحة آدم هاي؟ هاي يعني إني أرجل منكَ ومن كل عيلتك الخايخين	
(e)	ت.ف.ح	مْتَفَّح	مْتَفَّح	m t a f f i 7	ADJ:MS	reddish and healthy	مُحَمَر وصِحّي	وجهها مْتَفّح وحليانة كثير اسم الله	
(f)	ت.ف.ح	تَفَّح	تَفَّح	t a f f a 7	VERB:P	turn reddish and healthy	يصبح مُحَمَر وصِحّي	تَفّح وجهها بعد الجيزة. لاحظتوا؟	
(g)	ت.ف.ح	تَفَّح	يِتَفّح	y t a f f i 7	VERB:I	turn reddish and healthy			
(h)	ت.ف.ح	تَفَّح	تَفّح	t a f f i 7	VERB:C	turn reddish and healthy			

Table 1: Eight entries from Maknuune that share the same root, and are paired with four distinct lemmas.

CAPHI++	CAPHI	CAPHI Transcription	CODA	CAPHI++ Transcription
Q	k q 2 g	k aa l / q aa l / 2 aa l / g aa l	قَال	Q aa l
D	d dh	d ii b # dh ii b	ذِيب	D ii b
J	j dj	r i j j aa l # r i dj dj aa l	رِجّال	r i J J aa l
Z	z dh	z a n b / dh a n b	ذَنْب	Z a n b
T	t th	t i m m / th i m m	ثِمّ	T i m m
S	s th	th a w r a / s a w r a	ثَوْرَة	S a w r a
Z.	z. dh.	2 a z. u n n / 2 a dh. u n n	أَظُنّ	2 a Z u n
D.	d. dh.	b ee d. / b ee dh.	بَيض	b ee D.
K	k tsh	k ee f / tsh ee f	كِيف	K ee f

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

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 قَلَم *qalam* (*/q a l a m/*, */k a l a m/*, */2 a l a m/*, */g 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 Fajjar 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 tagsets 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 Emirtai 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/NOUN (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 (multiword 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 Glosses, 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

Root	Lemma	Form	Transcription	POS:Features	English	MSA	Example
ب.ي.ت	بَيت	بَيت مَضْوِي	b ee t # m a D. w i	NOUN:PHRASE	the parents have many children, especially males		
ب.ي.ت	بَيت	بَيت مَلْيَان	b ee t # m a l y aa n	NOUN:PHRASE	the parents have many children		
ب.ي.ت	بَيت	بَيت رُمَّان	b ee t # r u m m aa n	NOUN:PHRASE	the parents have many children		
ب.ي.ت	بَيت	بَيت مْعَمّ	b ee t # m 3 a t t i m	NOUN:PHRASE	there are no children at all in the house # the parents did not give birth to any children		
ب.ي.ت	بَيت	بَيت خَرَاب	b ee t # kh a r aa b	NOUN:PHRASE	all of the children are females # there are no male children in the house		أخوك عادي مسخمط بَيتُه خَرَاب، الله ما طعمه ولاد
ب.ي.ت	بَيت	بَيت عَامِر	b ee t # 3 aa m i r	NOUN:PHRASE	a house that is full of gatherings and happy celebrations		
ب.ي.ت	بَيت	بَيت عَمْرَان	b ee t # 3 a m r aa n	NOUN:PHRASE	a house that is full of gatherings and happy celebrations		
ب.ي.ت	بَيت	بَيت أَجِر	b ee t # 2 a J i r	NOUN:PHRASE	funeral	جَنَازَة	عملوله بَيت أجِر مسكين؟
ب.ي.ت	بَيت	بِفْتَح بَيت	y i f t a 7 # b ee t	NOUN:PHRASE	pay for the necessaties and needs of a family		هذا الراتب يا يابا ما بِفتح بَيت بطولكم
ب.ي.ت	بَيت	سِتّ بَيت	s i t t # b ee t	NOUN:PHRASE	housewife # the wife who can cook and clean the house very well	ربّة منزِل	بِديش أتجوز وحدة موظفة. بدي إياها سِت بيت
ب.ي.ت	بَيت	بَيت الخَارِج	b ee t # 2 i l kh aa r i J	NOUN:PHRASE	bathroom	حمَّام	كُنّا نروح عشي اسمه بَيت الخارج ما بقى في حمامات زي هلا
ب.ي.ت	بَيت	بَيت المَيّ	b ee t # 2 i l m a. y y	NOUN:PHRASE	bathroom	حمَّام	ودّي اخوتك عبَيت المي
ب.ي.ت	بَيت	بَيت خَالْتِي	b ee t # kh aa l t i	NOUN:PHRASE	prison	سِجن	كان عندي مشوار هيك لبَيت خَالْتِي هههههه
ب.ي.ت	بَيت	بَيت المُونة	b ee t # i l m oo n e	NOUN:PHRASE	pantry	مخزن طعام	جيبيلي قنينة زيت جديدة من بَيت المُونة

Table 3: Examples of NC compounds in Maknuune for the lemma بَيت 'house'.

'baby camel' depending on its age: ذَلُول *ðaluwl /dh a l uu l/*, 'barely a few days old' and حْوَيِّر *H.way∼ir /7 w a y y i r/* 'around 14-15 months old'. Another complex example is the word تَلْجِيم *tal.jiym /t a l J ii m/* '[lit. harnessing or bridling]' which can refer also to 'reciting some verses from the Quran (Surat Al-Takweer, Ayat Al-Kursi or Surat Al-Hashr) on a razor or a thread and closing the razor or tying the thread and leaving them aside until a lost or missing riding animal has returned home.'

Finally, we provide usage examples for some entries, as well as grammatical or collection notes. Notes vary in type from *Collective Noun* and *Collected near Nablus*, to *Vulgar*.

5 Coverage Evaluation

We approximate the coverage of our lexicon by comparing it with the Curras corpus (Jarrar et al., 2016), the largest resource available for PAL.[9] Since Curras 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 Curras. Then, we provide some comparison between Maknuune and the lexicons of two morphological analyzers for MSA and EGY.

[9] Al-Haff et al. (2022) describe a revised version of that corpus, but it was not made available at the time of writing.

POS Type	Unique lemma:POS	Entries	Forms	Phrases
Nominals	10,871	16,258	13,449	2,809
Verbs	6,179	19,622	18,982	640
Other	254	324	263	61
Proper & Foreign	65	98	65	33
Total	**17,369**	**36,302**	**32,759**	**3,543**

Table 4: POS type and entry statistics in Maknuune.

5.1 Maknuune & Curras 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

	Statistics	Maknuune	Curras Lexicon
All Entries	All entries	36,302	16,067
	Unique lemma:POS	17,369	8,448
	Unique lemma:POSType	17,083	8,161
	Unique lemmas	16,821	7,925
	Unique POS	35	33
	Unique roots	3,703	
	Entries per root	9.6	
	Unique lemma:POS per root	4.5	
Inflected Forms	All inflected forms	32,759	16,067
	Unique POS:features	76	224
Phrases	All phrase entries	3,543	
	Unique POS	25	

Table 5: Side-by-side view of the statistics of both Maknuune and the lexicon extracted from Curras.

nominals. Phrasal entries account for 10% of all Maknuune entries, and close to three quarters of them are associated with nominals (63% of all lemmas).

The Curras Lexicon In order to compare Maknuune with Curras, we extract a lexicon, henceforth Curras Lexicon, out of the Curras corpus by uniquing 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 سكون *sukun* diacritic, stripping all the فتحة diacritics that appear before a ا*A*, converting the همزة وصل آ*Ä* to ا*A*, and stripping the كسرة (*i*) and فتحة (*a*) diacritics if they appear before ة *ħ*. 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$_{MSA}$), and an Egyptian Arabic lex-

icon (Calima$_{EGY}$).[10] The analysis reveals that 44% of Maknuune overlaps with Calima$_{MSA}$ at the lemma:POSType level (63% if all entries are dediacritized),[11] 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 آوَى /2 aa w a/ which is 'thread a needle' in PAL and 'shelter sb' in both MSA and PAL, بَطّ /*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 آخرة /*2 aa kh r e*/ 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 difference.

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., مَدّ *mad∼* 'extend;stretch' (in PAL and MSA), has different inflections for some parts of the paradigm: the 2nd person masculine plural is مَدَّيتوا *mad∼aytuwA* in PAL and مَدَدْتُم *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, AbdelRahim 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.

[10] 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.

[11] 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, iraqi 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, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. 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 Elihai. 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.

Moïn 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.

Simon Hopkins. 1995. sarār "pebbles" — A Canaanite Substrate Word in Palestinian Arabic. *Zeitschrift für arabische Linguistik*, (30):37–49.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Salam Khalifa, Sara Hassan, and Nizar Habash. 2017. A morphological analyzer for Gulf Arabic verbs. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Frank Rice and Majed Sa'id. 1979. *Eastern Arabic*. Georgetown University Press.

Majdi Sawalha, Faisal Alshargi, Abdallah AlShdaifat, Sane Yagi, and Mohammad A. Qudah. 2019. Construction and annotation of the Jordan comprehensive contemporary Arabic corpus (JCCA). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 148–157, Florence, Italy. Association for Computational Linguistics.

Ken-ichi Seto. 1999. Distinguishing metonymy from synecdoche. *Metonymy in language and thought*, 4:91–120.

June E Shoup. 1980. Phonological aspects of speech recognition. *Trends in Speech Recognition*, pages 125–138.

K. Stowasser and M. Ani. 2004. *A Dictionary of Syrian Arabic: English-Arabic*. G - Reference, Information and Interdisciplinary Subjects Series. Georgetown University Press.

Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash. 2018. An Arabic Morphological Analyzer and Generator with Copious Features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON)*, pages 140–150.

Raphael Talmon. 2004. 19th century Palestinian Arabic: the testimony of Western travellers. *Jerusalem studies in Arabic and Islam*, (29):210–280.

A. Younis and M. Aldrich. 2021. *Palestinian Colloquial Arabic Vocabulary*. Arabic Vocabulary. Lingualism.

A POS Type Mapping and Examples

POS Type	POS	Examples
Nominals	ADJ	أَسْوَد، سُودَا، سُود
	ADJ_COMP	أَكْبَر، أَصْغَر
	ADJ/NOUN	نِيقَة، خُرْخُشَة، عُبْرَة، خِيخَة
	NOUN	وَلَد، وْلَاد، زَلَمَة، زْلَام، لِبرَة
	NOUN_ACT	بقيت بالزمانات كَاتِب قصة قصيرة
	NOUN_PASS	كل شي مَكْتُوب عالكتاب تبعي؟
	NOUN_QUANT	بَعْض، نُصّ، كُلّ، أَغْلَب
Verbs	VERB	لِعِب، يِلْعَب، اِلْعَب
Proper	NOUN_PROP	نُور، عَائِشَة
Other	ABBREV	إِلَخ
	ADJ_NUM	أَوَّل، ثَانِي، ثَالِث، رَابِع
	ADV	هَون، هَيك، هْنَاك، هَلَّا
	ADV_INTERROG	شْلَونَك؟
	ADV_REL	وَين ما يروح بلاقيه بوجهي
	CONJ	كلنا قعدنا عنفس السفرة حَتَّى ولادهم الصغار
	CONJ_SUB	تعبت كتير لَمَّا وصلت الدار
	INTERJ	وَلّ، نَعَم، لَا
	NOUN_NUM	وَاحِد، إِثْنِين، ثَلَاثَة
	PART	طَبّ أنت هلَا شو خصَّك فيني
	PART_DET	ال
	PART_FOCUS	أَمَّا بخصوص عمتي هند، فاحنا مالازم نسكت
	PART_FUT	رَح، رَايِح
	PART_INTERROG	أنت خليلي شِي؟
	PART_NEG	مش، مُو
	PART_PROG	أنت مش عَم تعطيني فرصة أحكي
	PART_RESTRICT	كلهم مناح إِلَّا الكبيرة
	PART_VOC	يَا ولد !
	PREP	مِن، عَن، لَ، فِي
	PRON	أَنَا، إِحْنَا، إِنْتِن
	PRON_EXCLAM	مَا احلاها!
	PRON_DEM	هَذَا، هَذَولَاك، هَرْغُو، هَرْغِيتُو، هَرْغَيتَا
	PRON_INTERROG	كِيف، مَتَى، وَين، لَيش، اَيش
	PRON_REL	اللِّي، أَيّ
	VERB_NOM	إِصْحَى، أُوعَى
	VERB_PSEUDO	أَكِن، رَيت، كَأَنّ

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