Curras + Baladi: Towards a Levantine Corpus

Karim El Haff, Mustafa Jarrar*, Tymaa Hammouda, Fadi Zaraket
University of Strasbourg, Birzeit University, American University of Beirut
karim.el-haff@etu.unistra.fr, mjarrar@birzeit.edu, 1171779@student.birzeit.edu, fz11@aub.edu.lb

Abstract

This paper presents two-fold contributions: a full revision of the Palestinian morphologically annotated corpus (Curras), and a newly annotated Lebanese corpus (Baladi). Both corpora can be used as a more general Levantine corpus. Baladi consists of around 9.6K morphologically annotated tokens. Each token was manually annotated with several morphological features and using LDC's SAMA lemmas and tags. The inter-annotator evaluation on most features illustrates 78.5% Kappa and 90.1% F1-Score. Curras was revised by refining all annotations for accuracy, normalization and unification of POS tags, and linking with SAMA lemmas. This revision was also important to ensure that both corpora are compatible and can help to bridge the nuanced linguistic gaps that exist between the two highly mutually intelligible dialects. Both corpora are publicly available through a web portal.

Keywords: Arabic morphology, Annotated Corpus, Arabic Dialect, Levantine, Palestinian Arabic, Lebanese Arabic

1. Introduction

The processing of the Arabic language is a complex field of research. This is due to many factors, including the complex and rich morphology of Arabic, its high degree of ambiguity, and the presence of several regional varieties that need to be processed while taking into account their unique characteristics. When its dialects are taken into account, this language pushes the limits of NLP to find solutions to problems posed by its inherent nature. It is a diglossic language; the standard language is used in formal settings and in education and is quite different from the vernacular languages spoken in the different regions and influenced by older languages that were historically spoken in those regions. Indeed, Arabic speakers use those local varieties in day-to-day communication. We can distinguish several families of dialects: Moroccan, Egyptian, Sudanese, Levantine, Iraqi and Khaliji (Gulf). Arabic dialects tend to diverge from Modern Standard Arabic (MSA) in terms of phonetics, morphology, syntax and vocabulary.

Arabic content was mainly written in MSA. Recently, dialectal content has been increasing massively, especially on social media. MSA is considered among the under-resourced languages by the NLP community (Darwish et al., 2021). Dialectal Arabic (DA) is even less resourced. The resource gap between MSA and the dialects implies a large margin of error when MSA tools are used against dialectal content (Zbib et al., 2012). Thus, it is important to build resources and tools to identify dialects in context and to treat Arabic content based on its *unique* dialectal identity.

In this research, we focus on the Lebanese variety of Levantine Arabic, which is used in daily conversations and in the Lebanese media. It is spoken by about 6 million locals, and almost double that number in diaspora. The paper presents a morphologically annotated corpus

for Lebanese. The development of the corpus uses texts covering a wide spectrum of subjects and registers. The corpus is designed to be compatible with, and leverage, Curras (Jarrar et al., 2017), the Palestinian corpus with morphological annotations. In this way, both corpora can be used as a more general Levantine corpus, especially that the Palestinian dialect represents Southern Levantine and that Lebanese represents Northern Levantine varieties. In addition to providing new Lebanese corpus annotations, we have also revised Curras annotations to ensure compatibility with the LDC's SAMA tags and lemmas (Maamouri et al., 2010).

In this paper, we present two-fold contributions:

1. Baladi, a Lebanese morphologically annotated corpus, which consists of 9.6K tokens. Each token was manually annotated with prefixes, suffixes, stem, POS tags, MSA and DA lemmatization, English gloss, in addition to other features such as gender, number, aspect, and person. The corpus was annotated mainly using LDC's SAMA lemmas and tags. The inter-annotator evaluation on most features illustrates 87% agreement using the Cohen's Kappa score (McHugh, 2015).

2. Revision of Curras, by refining all annotations for accuracy, normalization and unification of POS tags, and linking with SAMA lemmas. This revision was also important to ensure that both corpora are compatible and can together form a more general Levantine corpus.

Both, Curras and Baladi, are publicly available online[1].

The rest of the paper is organized as follows. We overview related work in Section 2. Section 3 describes the Lebanese dialect. In Section 4, we present Corpus Baladi. In Section 5 we present the annotation process and guidelines. Section 6 presents the evaluation

---

* Corresponding author.

[1] https://portal.sina.birzeit.edu/curras

and the inter-annotator agreement. In Section 7, we present the revisions we introduce to Curras. Section 8 discusses how we managed to transform Curras into a more Levantine corpus. Finally, Section 9 concludes.

2. Related Work

This section reviews efforts to create annotated corpora for Arabic dialects as well as for MSA.

2.1. MSA Resources

The Penn Arabic Treebank (PATB) (Maamouri et al., 2005) by the Linguistic Data Consortium (LDC) is central to the development of several MSA resources. It enriches newswire text in MSA collected from several news outlets with tokenization, segmentation, lemma, POS and gloss tags annotations along with syntactic trees. PATB uses the morphological tags as defined by the BAMA morphological analyzer (Buckwalter, 2004), which provides vocalized solutions, unique lemmas, prefixes, suffixes, stems, POS tags, and English gloss terms. SAMA (Maamouri et al., 2010) is a substantial improvement and refinement on BAMA as it extends its lexicon and provides several analysis refinements. The Prague Arabic Dependency Tree bank (Hajič et al., 2004) enriched the literature with functional linguistic annotations which in turn lead to the emergence of ElixirFM (Smrž, 2007) The Arabic lexicographic database at Birzeit University (Jarrar and Amayreh, 2019; Alhafi et al., 2019) provides a large set of MSA lemmas, word forms, and morphological features, which are linked with the Arabic Ontology (Jarrar, 2021; Jarrar, 2011) using the W3C LEMON model (Jarrar et al., 2019)

2.2. Dialectal Resources

The Levantine Arabic Treebank (Maamouri et al., 2006) featured the Jordanian Arabic dialect. Curras (Jarrar et al., 2017; Jarrar et al., 2014) is a more recent Levantine corpus featuring the Palestinian dialect. Large number of textual entries were collected from Facebook, Twitter and scripts of the Palestinian series "Watan Aa Watar".Each word in the corpus was then manually annotated with a set of morphological attributes. The corpus contains 56K tokens.

Earlier, the CALLHOME Egyptian Arabic corpus (Canavan et al., 1997) consisted of transcripts of telephone conversations in Egyptian. CALIMA (Habash et al., 2012) extended ECAL (Kilany et al., 2002) which build on CALLHOME to provide morphological analysis functionality of the Egyptian dialect. The CO-LABA project (Diab et al., 2010) collected resources in dialectal Arabic (mainly in Egyptian and Levantine) from the collection of online blogs. The effort eventually lead to constructing the Egyptian Tree Bank (ARZATB) (Maamouri et al., 2014). Curras and ARZATB were leveraged as case studies for morphological analysis and disambiguation (Eskander et al., 2016). YADAC (Al-Sabbagh and Girju, 2012) focuses

also on the Egyptian dialect identification and provides a mutli-genre approach. It is a collection of web blogs, micro blogs, and several Egyptian content discussion forums. MADAR (Bouamor et al., 2014) is an ongoing multi-dialect corpora covering 26 different cities and their corresponding dialects. Other efforts cover Emirati (Ntelitheos and Idrissi, 2017; Khalifa et al., 2018), Tunisian and Algerian (Zribi et al., 2015; Harrat et al., 2014), and Yemeni and Moroccan (Al-Shargi et al., 2016).

Our proposed contributions in this paper is to enrich Curras by (1) providing a Lebanese Levantine extension and by (2) refining and revising Curras entries to better accommodate the general Levantine dialect.

3. Lebanese and Levantine Dialects

The Levantine family of dialects can be linguistically split into Northern Levantine including the Lebanese and Syrian dialects, and Southern Levantine including Palestinian and Jordanian. During the spread of Arabic from the seventh century onwards, the Levant was a region that mainly spoke Western Aramaic (Skaf, 2015). Aramaic is a Semitic language continuum spoken mainly during antiquity throughout the Levantine region and it served as a lingua franca then. Aramaic survives today through modern dialects such as Turoyo Syriac and Western Neo-Aramaic spoken in parts of Syria. It also survives more subtly in the noticeable substratum underlying Levantine dialects that differ from MSA on several common linguistic specifities such as phonology, syntax, morphology and lexicon. This motivates using dialect specific annotations to annotate Levantine dialects. In the sequel, we briefly review the differentiating factor between Levantine dialects and MSA.

3.1. Phonological differences

Like other Semitic languages, Aramaic and its varieties were written with a 22-letter alphabet (Abjad). When Arabic was spread to the Levant, the Christian populations of the region began to transcribe the Arabic language using this consonantal alphabet, a tradition of Syriac writing known as "Garshouni" (Briquel Chatonnet, 2005). Due to the lack of some letters compared to the Arabic alphabet which contains 28 letters, adaptations were made in the Garshouni script, and some Syriac graphemes can represent several phonemes of Arabic, especially among the emphatic letters. Indeed, certain Arabic phonemes were not widely used by Levantine populations, even to this day; a speaker of Lebanese today tends to de-emphasize emphatic letters in Arabic words as for example in the case of مظلوم (abused) which is pronounced *mẓlwm* in MSA and *mzlwm* in Lebanese. Another example may be words containing ث *ṯ* such as ثعلب (fox) that is pronounced *tʿlb* or *sʿlb* across the Levant. As phonology differs in many situations for Levantine dialect speakers, spelling can vary greatly and can pose a challenge to the processing of those dialects when written.

3.2. Syntactical differences

A common usage for sentence structure in MSA is the verb-subject-object (VSO) structure. Sentences tend to start with a verb followed by its subject and then its object. Other structure configurations tend to be less frequent. On the other hand, in Levantine dialects, this structure is more flexible as the verb and subject have a natural flow of interchangeable positions.

MSA (VSO):	أكل الولد التفاحة
LEVANTINE (VSO):	أكل الولد التفاحة
LEVANTINE (SVO):	الولد أكل التفاحة
In English (SVO):	The child ate the apple

3.3. Morphological differences

Levantine inherits templatic morphology where affixes play an important role from its Semitic roots. Major morphological differences exist when compared to MSA. One of them is the loss of case markings in Levantine. Additionally, there are Levantine-specific morphemes that do not exist in MSA such as عم *m* ; the present continuous mark that precedes imperfect verbs to indicate the continuity of the action. Its absence in Levantine indicates that the action is a general truth: أنا عم باكل (I am eating) أنا باكل (I eat). MSA has no such entity. Context alone indicates whether the action is continuous or a general truth; أنا آكل can mean both "I am eating at the moment" or "I factually eat". Other morphemes include رح *rḥ* and ح *ḥ* that are the future markers in Levantine dialects as opposed to MSA's س *s* and سوف *swf* . Furthermore, the progressive Levantine particle ب *b* (as in باكل *bākl*) that is used to indicate imperfective verbs does not exist in MSA.

3.4. Differences in lexicon

It is also important to notice that the Levantine lexicon is rich with old Aramaic words due to its pre-Arab heritage, as well as foreign loan words due to the Levant's location as a frequent passage of many civilisations.

4. Corpus Collection

We manually collected texts written in Lebanese from sources such as Facebook posts, blog posts and traditional poems. We collected a total of 9.6K tokens spanning over 424 sentences. We merged them all into a single text file, with an average of 22 words per sentence. The corpus was chosen based on a critical judgment to include several registers of Lebanese speech, hence the choice to include folk poems, and satirical texts from social networks and blog articles. We avoided text written in Arabizi (Arabic written using proprietary Latin letters) as this is not the goal of our corpus at this phase. As the size of the corpus is relatively small, we performed data collection manually through the retrieval of the transcripts of traditional Lebanese poems زجل *zğl* by local poet Jihad Assi, satirical Facebook posts written in the vernacular Lebanese dialect by Mohamad Jaber as well as some blog posts written in the Lebanese dialect (Bel-Lebneene blog).

We did not preprocess the text and kept the raw form. As such, we did not perform any unification of letter variations, removal of diacritics, or correction of typos. We based this on the selective quality of our corpora. We then tokenized the raw text. This produced a table with three columns (sentence ID, token ID, and token text). The table was represented in a modern shareable spread sheet tool where each token and its annotations stood on its own separate row. The annotators introduced annotations in separate columns each designated for a specific feature or tag.

5. Annotation Methodology

Four linguists carried out the annotation process over a period of ten months. We used AnnoSheet to carry out the manual annotations. AnnoSheet is a Google Sheet that we empowered by developing and adding advanced JavaScript methods to (1) assist the annotation process and (2) validate the proposed annotations. For each row in the AnnoSheet, we have 16 columns: *sentence ID, Token ID, token, CODA, prefix, stem, suffix, POS, MSA lemma, dialect lemma, gloss, person, aspect, gender, and number*. The annotation guidelines for each of these columns are described in the following subsections.

To speed up the annotation process, we uploaded the revised version of Curras annotations (See section 7) into another spreadsheet and allowed the annotator to look up candidate annotations from Curras. The JavaScript lookup method searches Curras and returns the top matching results. The annotator can then select one of the results, and edit the corresponding fields if needed. The annotator also has the option to fill in the annotations directly. To guide and control the quality of the annotations, we implemented several validation triggers in JavaScript to highlight potential mistakes. In addition, for each cell in AnnoSheet, we implemented a customized list, from which the annotator can select values based on the column. The lists are dynamic and they are populated with values that depend on the values in other cells of the same row.

We, additionally, developed a Google Colab application to validate all annotations in the AnnoSheet and to flag cells that may require corrections. The validation ran once daily. In this way, the annotators were able to annotate each word, in context, by re-using annotations from Curras, fully or partially, or by entering new annotations. As a Google Sheet, AnnoSheet allowed the annotators to also annotate the corpus cooperatively, write feedback to each other, and skip tokens they are not certain about. Table 1 illustrates a sentence in Lebanese and its full annotations.

5.1. Annotation Guidelines

This section presents the annotation guidelines for each of the different annotations tasks.

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح شوف ظل الدنيا لوما بقيلي عيون

Token	CODA	Prefix(es)	Stem	Sufix(es)	POS	Person	Aspect	Gender	Number	DA Lamma	MSA Lemma	Gloss
بروحي	بروحي	ب/PREP	روح/NOUN	ي/POSS_PRON_1S	NOUN	-	-	m	s	رُوح 1	رُوح 1	spirit
ساكنة	ساكنه		ساكن/ADJ	ة/NSUFF_FEM_SG	ADJ	-	-	f	s	ساكن 1	ساكن 1	residing
ممنون	ممنون		ممنون/ADJ		ADJ	-	-	m	s	مَمنُون 1	مَمنُون 1	grateful
عيونك	عيونك		عيون/NOUN	ك/POSS_PRON_2FS	NOUN	-	-	m	p	عَين 1	عَين 1	eye
عيوني	عيوني		عيون/NOUN	ي/POSS_PRON_1S	NOUN	-	-	m	p	عَين 1	عَين 1	eye
وانا	وانا	و/CONJ	انا/PRON_1S		PRON	1	-	m	s	أنا 1	أنا 1	I
عم	عم		عم/PROG_PART		PROG_PART	-	i	-	-	عَم 0	ظَلّ 1	continue
آخذ	آخذ	أ/IV1S	خذ/IV		VERB	1	i	m	s	أخَذ 1	أخَذ 1	take
المونة	المونة	ال/DET	مون/NOUN	ة/NSUFF_FEM_SG	NOUN	-	-	f	s	مُونة 1	مُونة 1	provisions
رح	رح		رح/FUT_PART		FUT_PART	-	-	-	-	رَح 0	سَوؤ 1	will
ظل	ظل		ظل/IV		VERB	1	i	m	s	ظَلّ 1	ظَلّ 1	remain
شوف	شوف		شوف/IV		VERB	1	i	m	s	شاف 0	نَظَر 1	looking
الدني	الدنيا	ال/DET	الدنيا/NOUN		NOUN	-	-	f	s	دُنيا 2	دُنيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما 0	لَؤلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain
عيون	عيون		عيون/NOUN		NOUN	-	-	m	p	عَين 1	عَين 1	eye

Table 1: A sentence in Lebanese with its full annotations. Translation: "My eyes are grateful to yours who dwell in my soul as I make my provisions and I shall still, even without my eyes, see the world whole."

CODA

The CODA tag (التهجِّة الاصطلاحية) of a token signifies the "correct" spelling of the token. Instead of annotating the exact token in the corpus, the idea is to unify the different spelling variations of the same word into one CODA spelling, and annotate this CODA. Due to the lack of standardized orthographic spelling rules for Arabic dialects, people tend to write words as they pronounce them; thus, the same word might be written in different ways. In fact, the same person may write the same word in different ways in the same sentence. For example, consider the word meaning 'a lot': which can be written as كثير *ktyr* and كتير *ktyr* . The second letter in the first word corresponds to the sound θ. The correct spelling in MSA is the first word with ث. However, the ث [th] sound is rarely pronounced in Lebanese and is often replaced by the ت [t] sound or the [s] sound. Similarly, the words بألكن *bʔalkn* and بقلكون *bqlkwn* are different orthographic variations of the same word, which means "I tell you". In this case, we write both in the the same CODA spelling بقلكن *bqlkn* . See more examples in Table 2.

Token		CODA	
بألكن *bʔalkn*	→	بقلكن *bqlkn*	
بقلكون *bqlkwn*	→	بقلكن *bqlkn*	
طريء *ṭryʔ*	→	طريق *ṭryq*	
هايدي *hāydy*	→	هيدي *hydy*	
عيونّ *ʕywnn*	→	عيونٌ *ʕywnun*	

Table 2: Example of words and their CODA spelling.

Since our goal in this corpus is to transform Curras to be a more general Levantine corpus, we chose to adopt the Palestinian CODA guidelines (Habash et al., 2015) for the Lebanese dialect. We made some modifications and simplifications to be adapted to cover more Levantine regionalisms, as will be discussed in the next sections.

It is notable to add that some slight spelling differences exist between Lebanese and Palestinian as the former is a northern Levantine variety while the latter is southern and regional differences exist. The most common examples of this lie in demonstrative pronouns where Palestinian tends to use more emphatic sounds than Lebanese; a masculine "this" is said هاظا *hāẓā* or هادا *hādā* in Palestinian, while هَيدا *haydā* in Lebanese. Another example is the use of م *m* in Palestinian to indicate a third person plural where Lebanese uses a ن *n* : "your house" is بيتكم *bytkm* in Palestinian and بيتكن *bytkn* in Lebanese. The differences in spelling are due to the differences in pronunciation across the Levant and have no effect over the total mutual intelligibility of the dialects and thus a potential standardized spelling for Northern and Southern Levantine can be seen as the slight differences between British English and American English spelling systems.

MSA Lemma

The MSA Lemma (الدخلة المعجمية الفصحى) is the MSA lemma of the token. We restricted the choices of MSA lemmas to SAMA lemmas. The AnnoSheet allows the annotator to search the SAMA database and select the target lemma. For tokens that are not derived from an

MSA lemma, like بدي *bdy* , we chose the closest SAMA lemma (e.g., أراد_1). In case, no matching MSA lemmas are found in the SAMA database, the annotator is allowed to look up lemmas from Birzeit's lexicographic database (Jarrar and Amayreh, 2019), which are linked with the Arabic Ontology (Jarrar, 2021) and represented in the w3C Lemon model (Jarrar et al., 2019). the annotator may also introduce a new MSA lemma, however, new lemmas are marked with "_0", such as (يوغا_0) or (هيفاء_0). Similar to SAMA lemmas, noun lemmas should be in the masculine singular form. Plural and feminine are acceptable in case there is no masculine singular. Verb lemmas should be in the past masculine singular 3rd person form.

Dialect Lemma

The dialectal lemma (المدخل المعجمية العامية) signifies the semantic value of the token as a lexicon entry. Similar to the MSA lemma, each token in the corpus is tagged with its DA lemma. If a token stems from MSA, then its MSA and DA lemmas are the same. For example, the dialect token بقلك *bqlk* , which means "I tell you", has the same MSA and DA lemma قال_1. Some Levantine lexicon instances differ from MSA and need their own dialectal lemmas. These lemmas potentially do not exist in an ordinary MSA dictionary, due to their likely origin in other languages, notably Aramaic. As an example, the typical Levantine words used to say 'inside' and 'outside' are جوّا *ǧwā* and برّا *brā* , respectively. These two words are different in MSA: 'inside' is داخل *dāḫl* and 'outside' is خارج *ḫārǧ* . Table 1 illustrates more examples of Levantine lemmas that are not in MSA, such as عَم *ʿam* , رَح *raḥ* , شاف *šāf* , لُوما *luwmā* .

Gloss

The gloss (المعنى بالانجليزية) is the meaning of the lemma in English. We restrict the glosses to be SAMA glosses if a SAMA lemma is used, or to Curras if available, otherwise we provide it in the same way.

Stem

The stem (الساق) is the base word after removing suffixes and prefixes from the token. We follow the ⟨Stem/POS⟩ tagging schema used in Curras and SAMA, where the stem and the POS are separated by '/'. The POS is limited to the exact stem POS tagset found in SAMA.

Prefixes and Suffixes

We follow the prefixes (السوابق) and suffixes (االواحق) tagging schema used in Curras and SAMA: {⟨Prefix1/POS⟩ + ⟨Prefix2/POS⟩...} and {⟨Suffix1/POS⟩ + ⟨Suffix2/POS⟩...}. As shown in Table 1 the prefix بـ in the word بروحي *brwḥy* is the preposition بـ/PREP. Multiple prefixes are combined with "+". For example, the three prefixes in the word وبالقلب are: و/CONJ+بـ/PREP+الـ/DET. Suffixes are written in the same way. For example, the suffixes in the word قلتلهن *qltlhn* are: تـ/PVSUFF_SUBJ:1S+ لـ/PREP+هن/PRON_3FP. Prefixes and suffixes are

critical when dealing with dialects. This is because the morphological difference between dialects and MSA words is mostly due to different combinations of prefixes and suffixes. Dialects use additional types of prefixes and suffixes that are not used in MSA. The prefix بـ in بيضلن, prefix هـ in هالعيون, or the suffix ش in بعرفش, are examples of affixes that are commonly used in Levantine dialects but are not part of the MSA morphology. To control the quality of our annotations of affixes (i.e., prefixes and suffixes), we extracted the set of all combinations of affixes in Curras and verified them manually (See section 7). This set along with the SAMA combinations of affixes were then uploaded to the AnnoSheet and used to limit the choices of the annotators.

Table 3 presents the set of the prefixes used in the revised version of Curras and uploaded into AnnoSheet to be used by the annotators. Prefixes in Palestinian and Lebanese are all in common but there are two exceptions. The ا in Palestinian can be used as an INTERROG_PART, like in اغنيها *āġnyhā* , which means "shall I sing it?". However, in such cases in Lebanese, the verb is conjugated in its imperfective first person form to express the same meaning by using با and is said باغنيها *bāġnyhā* . Additionally, all prepositions in Lebanese and Palestinian are the same, except for فـ which is used only in Palestinian. We would also like to note that there are two prefixes in the corpora that are used only in MSA forms, and not in Palestinian or Lebanese, which are the سـ/FUT_PART and the لـ/JUS_PART. They occur in both corpora due to code-switching as this is a common phenomenon in dialects.

Table 4 presents the set of suffixes used in the revised version of Curras. The majority of of the suffixes are common to both dialects. However, there seems to be one bold systematic difference between the two dialects and it concerns suffixes used to indicate a plural in the 2nd and 3rd person; كُن/هُن is used in Lebanese to always express a gender-neutral plural for the 2nd and 3rd person (e.g., بيتهُن *bythun* / بيتكُن *bytkun*) whereas its Palestinian counterpart uses كِن/هِن to mostly express a feminine 2nd and 3rd person plural (e.g., بيتهِن *bythin* / بيتكِن *bytkin*) aligning itself with MSA's بيتهُنَّ *bythunn* / بيتكُنَّ *bytkunn* . Nevertheless, the northern Palestinian variety is closer to that of Lebanese and uses كِن/هِن while remaining gender-neutral. Furthermore, Palestinian uses كو where Lebanese uses كُن for the 2nd person plural, respectively: بيتكو and بيتكُن. Palestinian also tends to use هم and كم, aligning itself with MSA's بيتهم and بيتكم, where Lebanese does not. These occurrences seem to be systematic and may be due to the fact that Lebanese is a Northern Levantine variety while Palestinian is Southern Levantine and such differences are bound to exist in the dialectal continuum, sometimes overlapping in border regions.

Part of Speech

The part of speech (POS) (قسم الكلام) concerns the grammatical category of the token. The annotators were

CONJ {و، ف}	وحكي، فحكي	PART_RESTRICT {الا}	والاهي
CONNEC_PART {ف}	فظلم	PROG_PART {عم}	بظلم
DEM_PRON {م}	هالبلد	REL_PRON {ما، ل}	ماتا، لين
DET {ال}	البنت	SUB_CONJ {و، ما، إن}	وتكفرا، ماتحبي، إنتا
EMPHATIC_PART {ت، ل}	لاوريكي، تاوريكي	VOC_PART {يا}	يامعلم
FUT_PART {ح، ه}	حريح، هريح	IV1S {ت، أ، ن}	بارمز، آخدها، بتذكر
INTERROG_PART{لا، ما، شو، اه، +ها}	ماانتاش، شوداك، اغتبها، ياغتبها	IV2S {ت}	بتدخل
JUS_PART {ل}	لنقول	IV2P {ت}	بتذكروا
NEG_PART {لا، ما، ش}	لا، ما، ماتنش، مشلام	IV3S {ت، ي}	بتحكي، بتحتحك
PREP {ل، ب، م، ع، علا، +كو، ت، ك، +نت}	لصاحياتها، بالبلد، مشان، عشان، علاحسان، بالبلد، تاحكيلك، كاكنجي، متشان	IV3P {ي}	بجيبوا
		IV1P {ن}	بنظم

Table 3: List of prefixes and their POS tags in MSA, Palestinian, and Lebanese, which we are used in both corpora. Palestinian-specific prefixes are marked with (*), and Lebanese with (+).

PREP {ل}	ستحلي	NSUFF_FEM_PL {ة}	اللبنانة
NEG_PART {ش، اش، وش}	بدش، مافكش، معيوش	PVSUFF_DO:3P {هن، +هُ}	هدهن، هدهُن
REL_ADV {ش}	للبد	PVSUFF_DO:2P {*كو، *كي، +كُن}	عرشكو، عرشكي،عرشكُن
POSS_PRON_1S {ي}	بدني	PVSUFF_DO:2FS {كي}	بشتكي
POSS_PRON_2P {*كو، *كي، +كُن}	بابكو، بابكي، بابكُن	PVSUFF_SUBJ:3P {وا}	طمعتيون
POSS_PRON_3S {ا}	بابتا	PVSUFF_SUBJ:3S {ت}	طالبتيني
POSS_PRON_3P {*هن، +هُن}	باتهن، باتهُن	IVSUFF_SUBJ:3S {ت}	بعجيضتي
POSS_PRON_2FS {كي}	بشكي	IVSUFF_DO:2P {*كو، *كي، +كُن}	بتلكو، بتلكي، تلنكُن
PRON_2FS {كي}	بابكي	IVSUFF_SUBJ:2P {*كو، *كي، +كُن}	بتلكو، بتلكي، بتلكُن
PRON_2P {*كو، *كي، +كُن}	بابكو، بابكي، بابكُن		

Table 4: List of suffixes and their POS tags in Palestinian and Lebanese corpora (MSA are excluded as they are many). Palestinian-specific suffixes are marked with (*), and Lebanese with (+).

limited to selecting the POS from the tagset used in SAMA.

Person

The person (الإسناد) refers to the person of the annotated word, if applicable. This can be the 1st, 2nd or 3rd person, which we represent by the numbers 1, 2 and 3.

Aspect

This column concerns, for verbs, their aspect (صيغة الفعل). We denote (i) for imperfective verbs (present tense, not completed), (p) for perfective verbs (completed, past tense), and (c) for imperative or command verbs.

Gender

The gender (الجنس) of the word, if applicable, which is (m) for masculine, (f) for feminine and (n) for neutral.

Number

The number (العدد) of the word, if applicable. which is (s) for singular, (p) for plural, or (d) for dual (to count two units).

6. Evaluation

In this section we evaluate the quality of the our annotations for the Lebanese corpus. We performed two evaluations: (*i*) Inter-annotator agreement using the Cohen's Kappa κ, and (*ii*) The F1-score between each annotator and an expert annotator. The results of the two

evaluations are summarized in Tables 5 and 6. To conduct the inter-annotator agreement, we randomly selected annotated sentences that together consist of 400 tokens, i.e., 4.2% of the corpus. We divided these sentences among our four annotators, such that each annotator re-annotates about 100 tokens that were annotated by another. We used these 400 new annotations to compute the Cohen's kappa κ agreement coefficient. The inter-annotation agreement per annotation feature was computed. Table 5 lists the name of the feature and the κ metric (Di Eugenio and Glass, 2004):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among annotators and p_e is the hypothetical expected agreement. In the second evaluation, an expert went over the 400 tokens and corrected the original annotations, if needed. We used these corrections to compute precision and recall where the main expert annotator was considered reference. The expert annotator performed the following correction actions per feature value:

- Approved a feature value annotation (increments tp : *true positives* for the feature value)

- Approved a missing feature value annotation (increments tn : *true negatives* for the feature value)

- Rejected a feature value annotation (increments fp : *false positives* for the feature value)

- Rejected a missing feature value annotation (increments fn : *false negatives* for the feature value)

The precision $\frac{tp}{tp+fp}$ reflects the ratio of the true positives over the sum of true positives and false positives. The recall $\frac{tp}{tp+fn}$ reflects the ratio of the true positives over the sum of true positives and false negatives. Table 6 reports the average precision and recall across all feature values for each feature. We also computed the F1-score based on the precision and recall as:

$$F1 - score = \frac{2 * precision * recall}{precision + recall}$$

The overall kappa k, precision, recall, F1-score for all features are calculated using the Python `sklearn.metrics` package. We present the average weighted by the support of each label for precision and recall. According to the interpretation of the κ score (McHugh, 2015), the aspect and the suffix features scored moderate agreement (between .4 and .6), the stem and the prefix features scored near perfect agreement (above .81), and the rest of the features scored substantial agreement (between .61 and .80). The precision and recall scores of the corrected items show values that concur with the κ coefficient.

These results reflect some areas of disagreement between the annotators. A notable example of this is with

Tag	Values	Agreement	Disagreement	Kappa
Stem	178	357	43	0.884
Prefix	41	380	20	0.860
Suffixes	55	358	42	0.738
POS	22	340	60	0.821
Person	3	359	41	0.629
Aspect	4	384	16	0.911
Gender	3	337	63	0.687
Number	4	347	53	0.741
Overall				0.785

Table 5: Cohen's Kappa coefficient for the inter-annotation agreement and the precision and recall metrics for the main expert corrections.

Feature	Precision	Recall	F1-Score
Stem	0.9036	0.8935	0.893
Prefixes	0.964	0.95	0.955
Suffixes	0.948	0.895	0.915
POS	0.898	0.85	0.853
Person	0.928	0.898	0.910
Aspect	0.974	0.96	0.967
Gender	0.845	0.843	0.844
Number	0.881	0.868	0.873
Overall	0.918	0.894	0.901

Table 6: The precision and recall metrics for the main expert corrections.

prepositions that have a pronoun attached to them such as معها where there should not be any gender or number assigned. In such an example, some annotators assigned the gender and number of the suffix ها to token مع. That was corrected to be a gender-less and numberless preposition. Other disagreements are present in some instances where the suffix ة that indicate the feminine gender are not annotated as a suffix but are merged with the stem of the word. Some differences in POS agreement are present for example in the case where gender-less and numberless adverbs are annotated as prepositions or interrogative adverbs (such as كيف) which is not a striking disagreement in itself.

7. Curras Revisions

In order to ensure compatibility with Curras annotations, tagsets, and lemmas, some revisions on Curras were necessary. Curras consists of 55,889 tokens. Each token was fully annotated with the morphological features that we adopted in section 5. Since the same token can be used in the same way (i.e., the same features) in different sentences, it is expected that the exact annotations will be repeated. However, we found that this is not always the case in Curras. For example, the same word يدوروا *ydwrwā* appeared in two different sentences in Curras, with the same meaning "searching for", but each time with a different MSA lemma: (بَحَث)

and (بحث); while it should be (بَحَث 1). The adverb بس *bs* was correctly annotated in all occurrences in Curras; however, in some cases, it was mistakenly assigned with gender; and in some cases, it was annotated with the noun POS. We also found some typos in the tagsets of the stems and affixes.

Our goal is to unify and normalize such variations, and then build a list of morphological solutions as clean as possible. We performed the following revision steps:

a. Tokenization and POS
We developed a POS parser that reads the prefixes, stem, and suffixes in a given solution (i.e., annotations of a token), and returns a validation flag. We carefully inspected solutions that were flagged for review. The POS parser validates the following: (*i*) no parsing errors in the prefixes, stem, and suffixes, (*ii*) the transliterations of the prefixes, stem, and suffixes in Buckwalter are correct, (*iii*) the concatenation of the prefixes, stem, and suffixes corresponds to that of the CODA, (*iv*) every prefix should be in the predefined set of prefixes, (*v*) every suffix should be in the predefined set of suffixes, and (*vi*) every stem POS should be in the SAMA POS tagset.

b. Lemmatization
Curras originally contained 8,560 unique lemmas. Although Curras was annotated using SAMA lemmas, some of Curras lemmas were incorrectly linked with SAMA lemmas. This was mostly because of partial diacritization of lemmas (e.g., بحث) or as the lemma subscript is ignored (e.g., بَحَث). Ignoring diacritics and subscripts makes the lemma ambiguous. Thus, we cannot know, for example, whether it is (بَحَث 1), (بَحّث 2), or (بَحْث 1). To disambiguate MSA lemmas in Curras and link them with SAMA lemmas, we developed a lemma disambiguator that takes the lemma, POS, and gloss, and tries to reduce the number of choices. In case one lemma is returned, it is then considered the correct SAMA lemma, otherwise undecided. We were able to disambiguate about 5,120 unique lemmas (i.e., 58%) in this way. The remaining undecided 3,560 lemmas were manually disambiguated and linked with SAMA. As a result, the unique number of MSA lemmas in Curras now is 7,313. These include 6,781 that are mapped with SAMA lemmas, and 432 MSA lemmas that are not found in SAMA. We marked the latter with "_0".

Validating and unifying dialect lemmas was straightforward. In case a dialect lemma has the same letters as the MSA lemma (i.e., ignoring diacritics and subscripts) then it is the same lemma. So, we replace the dialect lemmas with the MSA lemma. Otherwise, a manual verification is performed. As a result, the unique number of dialect lemmas in Curras is 8,510. These include 7,785 lemmas equivalent to MSA lemmas, and 1,012 dialect lemmas that have no corresponding MSA lemmas. We also marked the latter with "_0".

c. Other features

We applied some heuristics in cleaning the Person, Aspect, Gender, and Number features. For example, the Aspect and Person are assigned only to verbs, otherwise they should be "-". We compared the Gender and Number with the suffix tags which also indicate gender and number, and corrected mistakes manually when needed.

d. Generating Unique Solutions

We prepared a table with unique annotations from Curras, called the "Solutions" table. We reused these solutions to annotate the Lebanese corpus in order to maximize the compatibility between both corpora. To do this, we split Curras into two tables: Tokens and Solutions, with a solution identifier (id) to link them. The Tokens table contains only the token id, token, and solution id. In this manner, the tokens that have the exact same annotations are given the same solution id. The Solutions table contains all annotations after removing the exact redundancies, which consists of 16,244 solutions. We considered two solutions to be identical if they have the same: CODA, prefixes, stem, suffixes, DA lemma, MSA lemma, Person, Aspect, Gender, and Number. We uploaded the Solutions table into our AnnoSheet and enabled our annotators to look up and reuse annotations from the Solutions table, as described in Section 5. As a result of this effort, we envision that the revised version of Curras along with the additions from Baladi; the newly built Lebanese corpus, form a more Levantine dialect corpus.

8. Discussion: a more Levantine Corpus

In this section, we discuss how both Palestinian and Lebanese corpora can be used as one, more Levantine corpus. Not only they are annotated with the same tagsets as discussed earlier, but adding 9.6K annotated Lebanese tokens to the Palestinian corpus Curras has helped bridge the nuanced linguistic gaps that exist between the two highly mutually intelligible dialects. Those nuances, as discussed earlier in this paper, are notably present in the affixes (i.e., morphology). Indeed, some prefixes and suffixes are typically Palestinian and not habitually used in Lebanese and vice-versa. However, these differences are a few and the majority of affixes are common to both dialects (See the differences in section 5.1).

Additionally, Lebanese functional words have also been incorporated, solidifying our idea of a more Levantine corpus where the dialectal continuum is taken into account. In fact, the majority of the functional words are common to both dialects. Table 7 presents frequent functional words that are different in both dialects and the mapping between them. To summarize, both corpora consists of about 65.2K tokens, covering both Palestinian and Lebanese, annotated using the same guidelines.
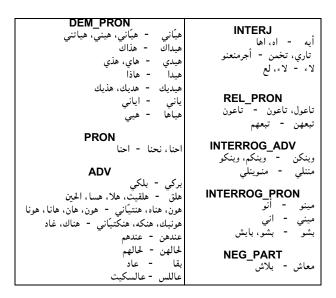
DEM_PRON		INTERJ	
هيّاني - هيّاني، هيني، هياتني		أيه - اه، اها	
هذاك - هيداك		تاري، تخمن - أجرمنعنو	
هاي، هذي - هيدي		لاء - لا، لع	
هاذا - هيدا		**REL_PRON**	
هديك، هذيك - هيديك		تاعول، تاعون - تاعون	
اياني - ياني		تبعهن - تبعهم	
هيي - هياها		**INTERROG_ADV**	
PRON		وينكن - وينكم، وينكو	
احنا، نحنا - احنا		مننلي - منوينلي	
ADV		**INTERROG_PRON**	
بركي - بلكي		مينو - أنو	
هلق - هلقيت، هلا، هسا، الحين		ميني - اني	
هون، هناه، هنتيّاني - هون، هان، هانا، هونا		بشو - بشو، بايش	
هونيك، هنكه، هنكتيّاني - هناك، غاد		**NEG_PART**	
عندهن - عندهم		معاش - بلاش	
لحالهن - لحالهم			
بقا - عاد			
عاللس - عالسكيت			

Table 7: Frequent functional words in Lebanese (right) and Palestinian (left).

9. Conclusions and Future Work

In this paper we presented the first morphologically annotated corpus for the Lebanese dialect 9.6K tokens. We also present a revised version of Curras, the Palestinian dialect corpus, about 55.9K tokens. We also described the various challenges we faced and measures we took to produce a compatible and more general Levantine Corpus, consisting of 55.9K tokens annotated with rich morphological and semantic information. Still, the evaluation of our annotators' performance shows a high degree of consistency and agreement. The Lebanese corpus is available for downloading and browsing online.

We plan to increase the size of our corpus to cover additional Levantine sub-dialects, especially those of other Levantine areas, most notably some of Syria's dialectal varieties. We also plan to use this corpus to develop morphological analyzers and word-sense disambiguation system for Levantine Arabic as we did for MSA (see (Al-Hajj and Jarrar, 2021a; Al-Hajj and Jarrar, 2021b)). Additionally, we plan to build on the Palestinian and Lebanese dialect lemmas to develop a Levantine-MSA-English Lexicon and extend it with synonyms (Jarrar et al., 2021). Both Curras and Baladi corpora are also being annotated with named-entities as part of the Wojood NER corpus see (Jarrar et al., 2022).

10. Acknowledgements

11. Bibliographical References

Al-Hajj, M. and Jarrar, M. (2021a). Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online, sep. INCOMA Ltd.

Al-Hajj, M. and Jarrar, M. (2021b). Lu-bzu at semeval-2021 task 2: Word2vec and lemma2vec performance in arabic word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 748–755, Online, aug. Association for Computational Linguistics.

Al-Sabbagh, R. and Girju, R. (2012). YADAC: Yet another dialectal Arabic corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1300–1306, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Alhafi, D., Deik, A., and Jarrar, M. (2019). Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE, November.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Briquel Chatonnet, F. (2005). De l'intérêt de l'étude du garshouni et des manuscrits écrits selon ce système. In *L'Orient Chrétien dans l'Empire musulman, en hommage au Professeur Gérard Troupeau*, Studia Arabica III, pages 463–475. Editions de Paris.

Buckwalter, T. (2004). Buckwalter arabic morphological analyzer version 2.0. *LDC2004L02*, dec.

Canavan, A., Zipperlen, G., and Graff, D. (1997). Callhome egyptian arabic speech. *LDC97S45*.

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S. R., El-Hajj, W., Jarrar, M., and Mubarak, H. (2021). A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81, April.

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

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. *LREC Workshop on Semitic Language Processing*, pages 66–74, 01.

Eskander, R., Habash, N., Rambow, O., and Pasha, A. (2016). Creating resources for dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3455–3465, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Habash, N., Eskander, R., and Hawwari, A. (2012). A morphological analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada, June. Association for Computational Linguistics.

Habash, N., Jarrar, M., Alrimawi, F., Akra, D., Zalmout, N., Bartolotti, E., and Arar, M. (2015). Palestinian arabic conventional orthography guidelines. Technical report, Birzeit University.

Hajič, J., Smrž, O., Petr, Z., Snaidauf, J., and Beška, E. (2004). Prague arabic dependency treebank: development in data and tools. *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, 01.

Harrat, S., Meftouh, K., Abbas, M., and Smaïli, K. (2014). Building resources for algerian arabic dialects. In *INTERSPEECH*.

Jarrar, M. and Amayreh, H. (2019). An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer, June.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pages 18–27. Association For Computational Linguistics, October.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2017). Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775, September.

Jarrar, M., Amayreh, H., and McCrae, J. P. (2019). Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings, May.

Jarrar, M., Karajah, E., Khalifa, M., and Shaalan, K. (2021). Extracting synonyms from bilingual dictionaries. In *Proceedings of the 11th International Global Wordnet Conference (GWC2021)*, pages 215–222. Global Wordnet Association, Jan.

Jarrar, M., Khalilia, M., and Ghanem, S. (2022). Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the International*

Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, June.

Jarrar, M. (2011). Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League, 7.

Jarrar, M. (2021). The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Al Kaabi, M. (2018). A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). Egyptian colloquial arabic lexicon. *LDC99L22*, jul.

Maamouri, M., Bies, A., Buckwalter, T., Jin, H., and Mekki, W. (2005). Arabic treebank: Part 3 (full corpus) v 2.0. *LDC2005T20*, June.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Ldc standard arabic morphological analyzer (sama) version 3.1. *LDC2010L01*, July.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

McHugh, M. L. (2015). Interrater reliability: the kappa statistic. *Biochemia medica*, 22.

Ntelitheos, D. and Idrissi, A. (2017). Language growth in child emirati arabic. *Perspectives on Arabic Linguistics XXIX*, 5:229–248.

Skaf, R. (2015). *Le morphème d= en araméen-syriaque : étude d'une polyfonctionalité à plusieurs échelles syntaxiques*. Theses, Université Sorbonne Paris Cité ; Università degli studi (Torino, Italia), November.

Smrž, O. (2007). ElixirFM – implementation of functional Arabic morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada, June. Association for Computational Linguistics.

Zribi, I., Ellouze, M., Belguith, L. H., and Blache, P. (2015). Spoken tunisian arabic corpus "stac": Transcription and annotation. *Res. Comput. Sci.*, 90:123–135.