The 2022 International Conference on Language Resources and Evaluation (LREC 2022)
Marseille, France
20-25/6/2022

Curras + Baladi: Towards a Levantine Corpus

Karim El Haff
Strasbourg University
France

Mustafa Jarrar
Birzeit University
Palestine

Tymaa Hammouda
Birzeit University
Palestine

Fadi Zaraket
American University of Beirut
Lebanon

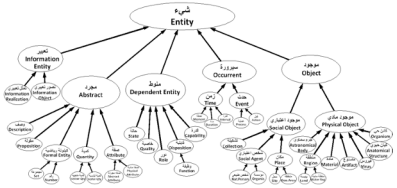
Lexical Resources at Birzeit University

Lexicographic Database

150 lexicons

Largest Arabic lexicographic database

Arabic Ontology/Wordnet



Formal Arabic Wordnet

with ontologically clean content

Dialect Corpora

Annotated corpora

each word is annotated with many morph features

NLP library

APIs

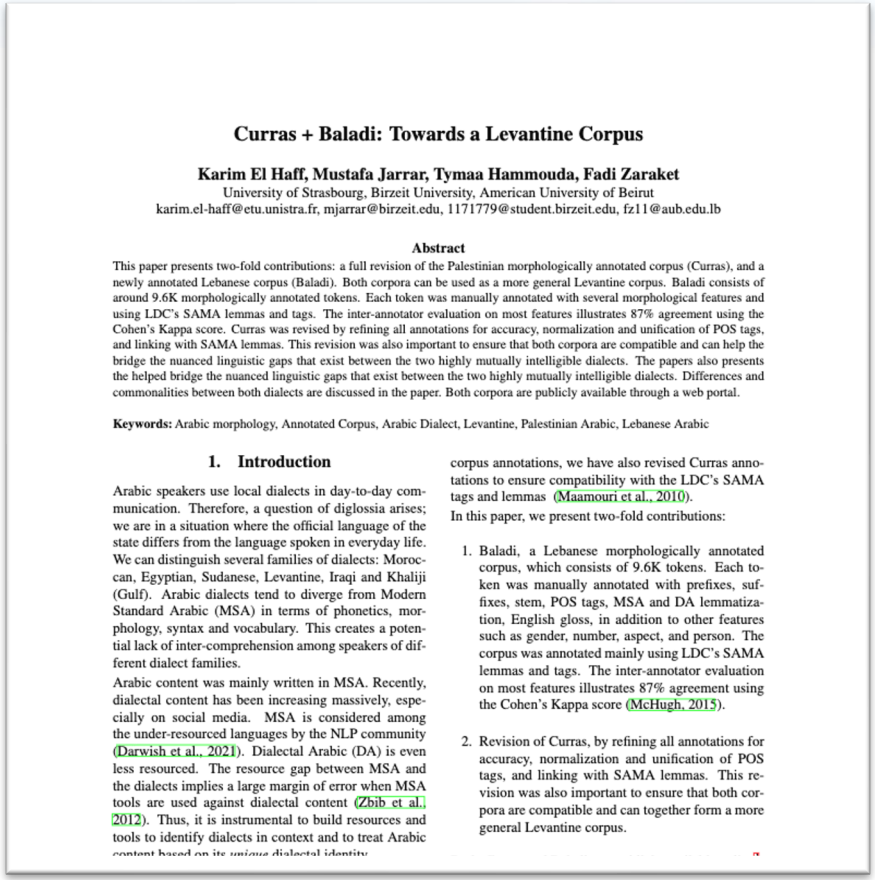
Linguistic Data, synonyms, tools, Nested named-entities, intents, …

WSD 84%

NER 88.4%

Big Linguistic Data Graph

https://ontology.birzeit.edu



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 87% agreement using the Cohen's Kappa 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 the bridge the nuanced linguistic gaps that exist between the two highly mutually intelligible dialects. The papers also presents the helped bridge the nuanced linguistic gaps that exist between the two highly mutually intelligible dialects. Differences and commonalities between both dialects are discussed in the paper. Both corpora are publicly available through a web portal.

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

1. Introduction

Arabic speakers use local dialects in day-to-day communication. Therefore, a question of diglossia arises; we are in a situation where the official language of the state differs from the language spoken in everyday life. 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. This creates a potential lack of inter-comprehension among speakers of different dialect families.

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 instrumental to build resources and tools to identify dialects in context and to treat Arabic content based on its *unique* dialectal identity

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).

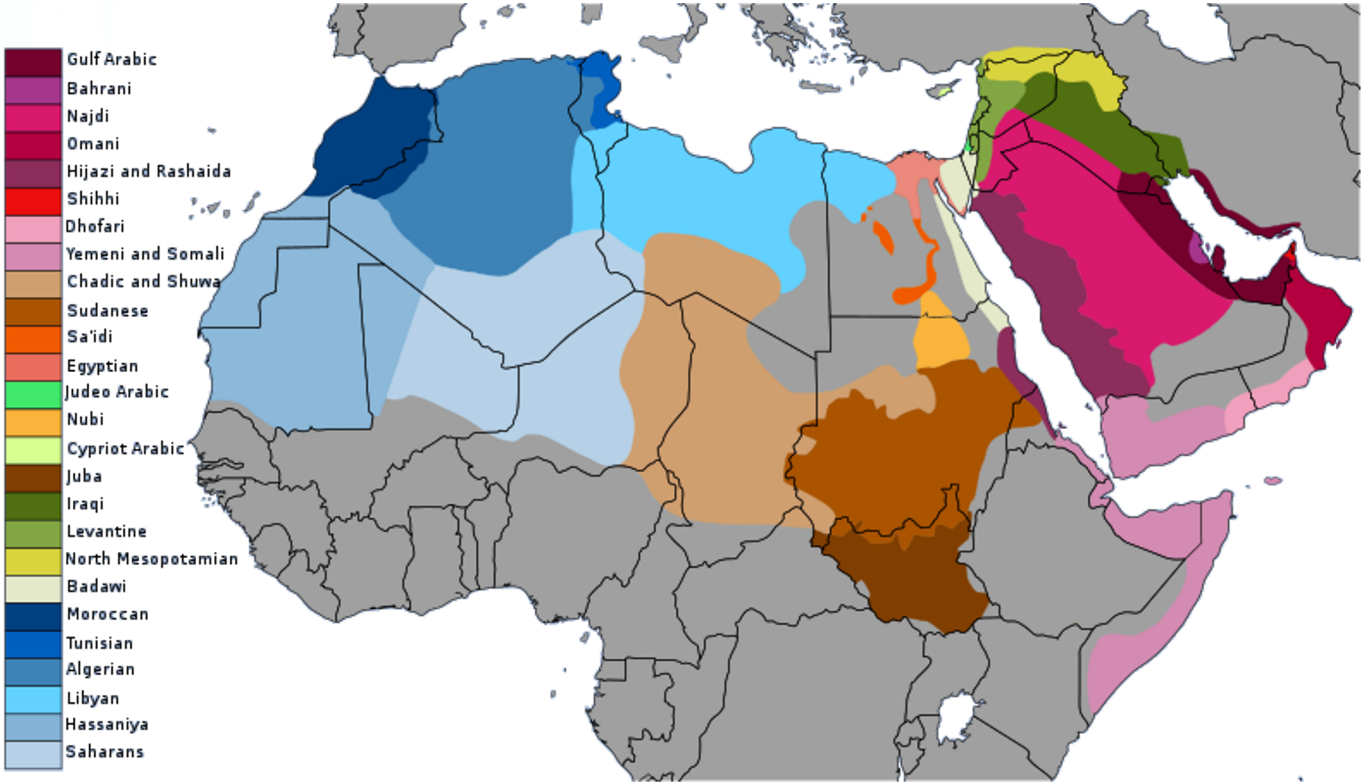
2. Revision of Curras, by refining all annotations for accuracy, normalization and unification of POS tags, and linking with SAMA lemmas. This revision was also important to ensure that both corpora are compatible and can together form a more general Levantine corpus.

El Haff, K., Jarrar, M., Hammouda, T., Zaraket, F., (2022). Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

The problem

Arabic is a low resources language

- Classical Arabic
- Modern Standard Arabic (more resources)
- **Arabic Dialects**



Our Contributions

Curras, a Palestinian morphologically annotated corpus (Revised)

Baladi, a Lebanese morphologically annotated corpus

Levantine Corpus

Publicly available:

https://portal.sina.birzeit.edu/curras

Baladi Lebanese Corpus

Corpus Collection:

From: Facebook posts, blog posts and traditional poems.
Size: 9.6K tokens, 424 sentences

Annotation Methodology

Tool: a *Smart* Google Sheet (**AnnoSheet**)
Four annotators, over 10 months
Features:

Token	CODA	Prefix(es)	Stem	Sufix(es)	POS	Person	Aspect	Gender	Number	DA Lamma	MSA Lemma	Gloss
المونې	المونة	ال/DET	مون/NOUN	ة/NSUFF_FEM_SG	NOUN	-	-	f	s	مونة 1	مونة 1	provisions

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM_SG	ADJ	-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ		ADJ	-	-	m	s	مَمْنُون 1	مَمْنُون 1	grateful
عيونك	عيونك		عيون/NOUN	ك/POSS_PRON_2FS	NOUN	-	-	m	p	عَيْن 1	عَيْن 1	eye
عيوني	عيوني		عيون/NOUN	ي/POSS_PRON_1S	NOUN	-	-	m	p	عَيْن 1	عَيْن 1	eye
وانا	وانا	و/CONJ	انا/PRON_1S		PRON	1	-	m	s	أَنا 1	أَنا 1	I
عم	عم		عم/PROG_PART		PROG_PART	-	i	-	-	عَم_0	ظَلّ 1	continue
آخذ	آخذ	آ/IV1S	خذ/IV		VERB	1	i	m	s	أَخَذْ 1	أَخَذْ 1	take
الموني	المونة	ال/DET	مون/NOUN	ة/NSUFF_FEM_SG	NOUN	-	-	f	s	مُونَة 1	مُونَة 1	provisions
رح	رح		رح/FUT_PART		FUT_PART	-	-	-	-	رَح_0	سَوْفَ 1	will
ضل	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	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقي/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقَ_0	بَقَ_0	remain

Annotation Guidelines

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

Token	CODA	Prefix(es)	Stem	Sufix(es)	POS	Person	Aspect	Gender	Number	DA Lamma	MSA Lemma	Gloss
بروحي	بروحي	ب/[illegible]EP	روح/NOUN	ي/POSS_PRON_1S	NOUN	-	-	m	s	رُوح 1	رُوح 1	spirit
ساكنه	ساكنة		[illegible]ADJ	ة/NSUFF_FEM_SG	ADJ	-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/AD[illegible]		ADJ			m		[illegible]	[illegible]	grateful
عيونك	عيونك		عيون/NOUN	[illegible]S_PR[illegible]								
عيوني	عيوني		عيون/NOUN	ي								
وانا	وانا	و/CONJ	انا/PRON_1S									
عم	عم		عم/PROG_PART									
آخذ	آخذ	آ/IV1S	خذ/IV									
المونة	المونة	ال/DET	مون/NOUN	ة/NSUFF_FE								
رح	رح		رح/FUT_PART									
ظل	ظل		ظل/IV									
شوف	شوف		شوف/IV		VERB	1	i	m	s	شاف_0	نَظَر 1	looking
الدني	الدنيا	ال/DET	الدنيا/NOUN		NOUN	-	-	f	s	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain

CODA (التهجئة الصحيحة)

- The “correct”/unified spelling of the token.
- Followed Curras CODA
- Examples:

Token		CODA
بألكن *b'alkn*	→	بقلكن *bqlkn*
بقلكون *bqlkwn*	→	بقلكن *bqlkn*
طريء *ṭry'*	→	طريق *ṭryq*
هايدي *hāydy*	→	هيدي *hydy*
عيونن *ʿywnn*	→	عيونٌ *ʿywnun*

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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	/POSS_PRON_1S	NOUN	-	-	m	s	رُوح1	رُوح 1	spirit
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_		-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ									
عيونك	عيونك		عيون/NOUN									
عيوني	عيوني		عيون/NOUN									
وانا	وانا	و/CONJ	انا/PRON_1S									
عم	عم		عم/PROG_PA									
آخذ	آخذ	آ/IV1S	خذ/IV									
الموني	المونة	ال/DET	مون/NOUN									
رح	رح		رح/FUT_PART									
ضل	ظل		ظل/IV									
شوف	شوف		شوف/IV									
الدني	الدنيا	ال/DET	الدنيا/NOUN									
لوما	لوما		لوما/CONJ									
بقيلي	بقيلي		بقي/IV									

Prefixes (سوابق)

- Based on SAMA prefixes tagset, which are:

CONJ { و ، ف }	وحكى، فحكى	PART_RESTRICT { الا }	والاهي
CONNEC_PART { ف }	فعلشان	PROG_PART { ب }	ببطلع
DEM_PRON { ه }	هالبلد	REL_PRON { ما ، ل }	ماشاء، لمين
DET { ال }	الغنية	SUB_CONJ { و ، ما ، إن }	وشكلوا، ماانتي، إنشاء
EMPHATIC_PART { ل ، ت }	لاوريكي، تاوريكي	VOC_PART { يا }	يامعلم
FUT_PART { ح ، سـ }	حيرح، سيريح	IV1S { إ ، آ ، ت }	بامزمز، آخدها ، بتذكر
INTERROG_PART{لا، ما، شو، *ا ، +با}	مالقاش ، شوبدك، اغنيها ، باغنيها	IV2S { ت }	بتدخل
JUS_PART { ل }	ليقول	IV2P { ت }	بتتذكروا
NEG_PART { لا ، ما ، مش }	ولاشي، مابقتش، مشلازم	IV3S { ت ، ي }	بتبكي، بيضحك
PREP { ل ، ب ، م ، ع ، عل ، *ف ، ت ، ك ، من }	لصاحباتها، بالبلد، مشان، عشان، علحصان فالبلد، تاحكيلك، كدكنجي، منشان	IV3P { ي }	بيجيبوا
		IV1P { ن }	بنطلع

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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_PH	OUN	-	-	m	s	رُوح1	رُوح 1	spirit
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM		-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ									
عيونك	عيونك		عيون/NOUN									
عيوني	عيوني		عيون/NOUN									
وانا	وانا	و/CONJ	انا/PRON_1S									
عم	عم		عم/PROG_PA									
آخذ	آخذ	آ/IV1S	خذ/IV									
المونة	المونة	ال/DET	مون/NOUN									
رح	رح		رح/FUT_PART									
ظل	ظل		ظل/IV									
شوف	شوف		شوف/IV									
الدنيا	الدنيا	ال/DET	الدنيا/NOUN									
لوما	لوما		لوما/CONJ									
بقيلي	بقيلي		بقي/IV									

suffixes (لواحق)

- Based on SAMA suffix tagset, which are:

صحلي	PREP { لـ}	اللبنانية	NSUFF_FEM_PL { ة}
بديش، ملقاش، معهوش	NEG_PART {ش ، اش ، وش}	هددهن، هددهُن	PVSUFF_DO:3P {*هِن، +هُن}
قديش	REL_ADV {يش}	عرفتكو، عرفتكن،عرفتكُن	PVSUFF_DO:2P { *كو، *كِن، +كُن}
بعدني	POSS_PRON_1S { ني}	شفناكي	PVSUFF_DO:2FS {كي}
بالكو، بالكِن ، بالكُن	POSS_PRON_2P {*كو، *كِن، +كُن}	طعميتوهن	PVSUFF_SUBJ:3P {تو}
تاعت	POSS_PRON_3S {ت}	طايقيتني	PVSUFF_SUBJ:3S {يت}
تاعتهِن، تاعتهُن	POSS_PRON_3P {*هِن، +هُن}	معجبيتني	IVSUFF_SUBJ:3S {يت}
عنيكي	POSS_PRON_2FS { كي}	بتقلكو، بتقلكِن، تقلكُن	IVSUFF_DO:2P { *كو، *كِن، +كُن}
عليكي	PRON_2FS { كي}	يديكو، يديكِن، يديكُن	IVSUFF_ SUBJ:2P { *كو، *كِن، +كُن}
نيالكو، نيالكِن، نيالكُن	PRON_2P { *كو، *كِن، +كُن}		

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM_		-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ									
عيونك	عيونك		عيون/NOUN	ك/POSS_PR								
عيوني	عيوني		عيون/NOUN	ي/POSS_PR								
وانا	وانا	و/CONJ	انا/PRON_1S									
عم	عم		عم/PROG_PART									
آخذ	آخذ	آ/IV1S	خذ/IV									
الموني	المونة	ال/DET	مون/NOUN	ة/NSUFF_FEM_SG	NOUN	-	-	f	s	مُونَة 1	مُونَة 1	provisions
رح	رح		رح/FUT_PART		FUT_PART	-	-	-	-	رَح_0	سَوْفَ 1	will
ضل	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	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain

Stem (الساق)

- ⟨Stem/POS⟩ tagging schema used in Curras
- SAMA POS tagset
- Reused Curras stems (as much as possible)

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM_SG		-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ		DJ	-	-	m	s	مَمْنُون 1	مَمْنُون 1	grateful
						-	-	m	p	عَيْن 1	عَيْن 1	eye
						-	-	m	p	عَيْن 1	عَيْن 1	eye
						1	-	m	s	أَنا 1	أَنا 1	I
					RT	-	i	-	-	عَم_0	ظَلّ 1	continue
						1	i	m	s	أَخَذْ 1	أَخَذْ 1	take
						-	-	f	s	مُونَة 1	مُونَة 1	provisions
					RT	-	-	-	-	رَح_0	سَوْفَ 1	will
ضل	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	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain

- Same SAMA tagsets: POS, Person, Aspect, Gender, Number

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM_SG	ADJ	-	-	f		ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ		ADJ	-	-		s	مَمْنُون 1	مَمْنُون 1	grateful
								m	p	عَيْن 1	عَيْن 1	eye
							-	m	p	عَيْن 1	عَيْن 1	eye
						1	-	m	s	أَنا 1	أَنا 1	I
					RT	-	i	-	-	عَم_0	ظَلّ 1	continue
						1	i	m	s	أَخَذْ 1	أَخَذْ 1	take
						-	-	f	s	مُونَة 1	مُونَة 1	provisions
					RT	-	-	-	-	رَح_0	سَوْفَ 1	will
ضل	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	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain

MSA Lemma (مدخلة معجمية فصحى)

- SAMA lemmas
- If no SAMA lemma, add one in the same way, e.g., (يوغا-0)

Annotation Guidelines

DA Sentence: بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف 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	[illegible]	رُوح 1	رُوح 1	spirit
ساكنه	ساكنة		ساكن/ADJ	ة/NSUFF_FEM_SG	ADJ	-	-	f	s	ساكِن 1	ساكِن 1	residing
ممنون	ممنون		ممنون/ADJ		ADJ	-		m	s	مَمْنُون 1	مَمْنُون 1	grateful
							-	m	p	عَيْن 1	عَيْن 1	eye
							-	m	p	عَيْن 1	عَيْن 1	eye
						1	-	m	s	أَنا 1	أَنا 1	I
					[illegible]RT	-	i	-	-	عَم_0	ظَلّ 1	continue
						1	i	m	s	أَخَذْ 1	أَخَذْ 1	take
						-	-	f	s	مُونَة 1	مُونَة 1	provisions
					[illegible]RT	-	-	-	-	رَح_0	سَوْفَ 1	will
ضل	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	دُنْيا 2	دُنْيا 2	world
لوما	لوما		لوما/CONJ		CONJ	-	-	-	-	لُوما_0	لَوْلا 1	except for
بقيلي	بقيلي		بقى/IV	لي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	بَقِيَ 1	بَقِيَ 1	remain

Dialect Lemma (مدخلة معجمية عامية)

- from MSA, then its MSA and DA lemmas are the same.
- If not, introduce a new DA lemma

Annotation Guidelines

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

INTERJ أيه - اه، اها تاري، تخمن - أجرمنعنو لاء - لاء، لع **REL_PRON** تاعول، تاعون - تاعون تبعهن - تبعهم **INTERROG_ADV** وينكن - وينكم، وينكو مننلي - منوينلي **INTERROG_PRON** مينو - أنو ميني - اني بشو - بشو، بايش **NEG_PART** معاش - بلاش	**DEM_PRON** هيّاني - هيّاني، هيني، هياتني هيداك - هذاك هيدي - هاي، هذي هيدا - هاذا هيديك - هديك، هذيك ياني - اياني هياها - هيي **PRON** احنا، نحنا - احنا **ADV** بركي - بلكي هلق - هلقيت، هلا، هسا، الحين هون، هناه، هنتيّاني - هون، هان، هانا، هونا هونيك، هنكه، هنكتيّاني - هناك، غاد عندهن - عندهم لحالهن - لحالهم بقا - عاد عاللس - عالسكيت

Baladi Evaluation

- Randomly selected annotated sentences that together consist of 400 tokens
- Each annotator re-annotated 100 tokens that were annotated by another

❖ Cohen's Kappa inter-annotation agreement

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

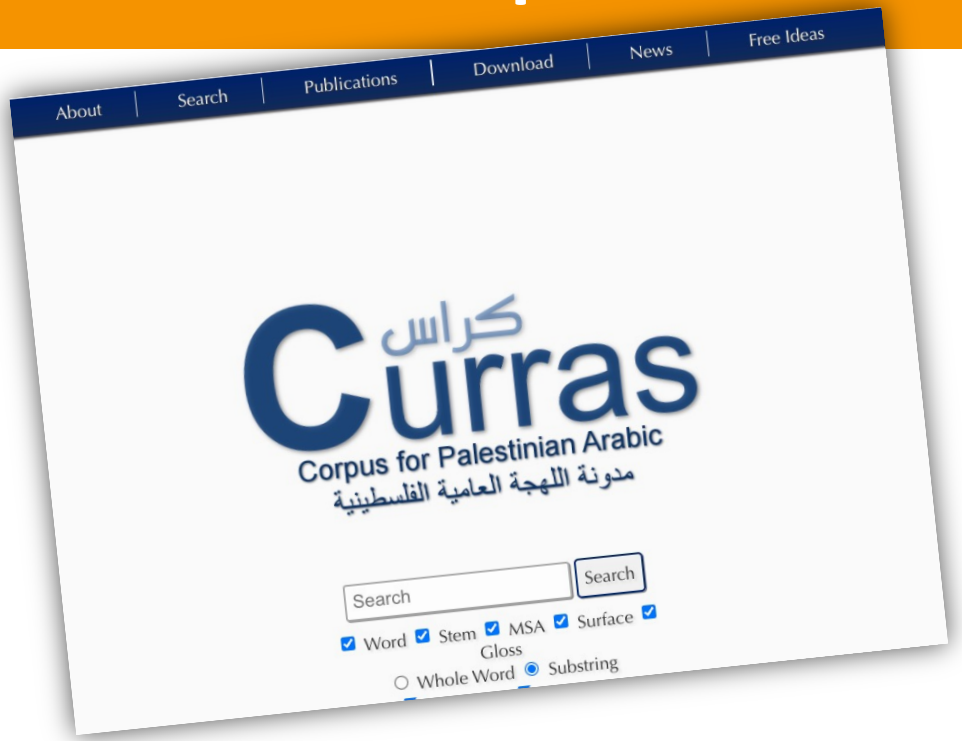
❖ F1-Score, precision and recall metrics

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

Curras – Palestinian Dialect Corpus

Originally

- 56K tokens
- morphologically annotated corpus



Jarrar, M., Habash, N., Alrimawi, F., Akra, D., & Zalmout, N. (2016). **Curras: An Annotated Corpus for the Palestinian Arabic Dialect**. Language Resources and Evaluation, 50(219), 1-31.

Jarrar, M., Habash, N., Akra, D. F., & Zalmout, N. (2014). **Building a corpus for Palestinian Arabic: a preliminary study**. In Proceedings – Arabic Natural Language Processing Workshop, at the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Association for Computational Linguistics. (pp. 18-27).

Curras Revisions

Almost all annotations are revised

1. Tokenization and POS

- No parsing errors in tokenization and POS
- Buckwalter transliterations
- CODA corresponds to prefixes + stem + suffixes
- every prefix should be in the predefined set of prefixes,
- every suffix should be in the predefined set of suffixes,
- every stem POS should be in the SAMA POS tagset.

Curras Revisions

Almost all annotations are revised

2. Lemmatization

Manually revised MSA and DA lemmas

- MSA lemmas are linked with SAMA lemmas, new lemmas are marked with "_0"
- DA lemmas are linked with MSA lemmas (if different)

Curras Revisions

Almost all annotations are revised

3. Other features

Person, Aspect, Gender, and Number are verified

Generating Unique Solutions

A minmum set of unique morphological solutions is generated and used to annotated the Lebanese corpus

Available Online

https://portal.sina.birzeit.edu/curras

Summary

Baladi: Lebanese morphologically annotated corpus (9.6K)

Curras (revised): Palestinian morphologically annotated corpus (56K)

= a more **Levantine Corpus** (65.6K tokens)

Differences and communities are discussed in the paper

Thank You

References