Curras + Baladi: Towards a Levantine Corpus

Karim El Haff[a], Mustafa Jarrar[b], Tymaa Hammouda[b], Fadi Zaraket[c]

[a] University of Strasbourg, [b] Birzeit University, [c] American University of Beirut

Contributions

- Revision of Curras[1]:
  - *Palestinian* morphological corpus
- Creation of Baladi:
  - Newly annotated *Lebanese* corpus
- A more Levantine corpus [2]:
  - Corpora compatibility, POS normalization and unification, and Lemma coherence
  - https://portal.sina.birzeit.edu/curras

Baladi: Lebanese corpus

- Collection of 424 sentences
  - 9.6K tokens
  - Social media and blog posts
  - Traditional Lebanese poems
- Methodology: smart Google sheets
- Features:
  - CODA, Prefix, Stem, POS, Person, Aspect, Gender, Number, Dialect Lemma, MSA Lemma, and Gloss

Lemma annotation guidelines

MSA Lemma مدخلة معجمية فصحى

- SAMA lemmas: If none, add one in the same way, e.g., يوغا-0

Dialect Lemma مدخلة معجمية عامية

- From MSA, then MSA and DA lemmas are the same. Otherwise, introduce a new DA lemma

Leb/Pal functional words

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

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

Evaluation

- Randomly selected annotated sentences
- A total of 400 unique tokens
  - Annotator overlap
  - Each annotator re-annotated 100 tokens
  - Each of the 400 tokens annotated twice

References

[1] M. J. et al., "Curras: An annotated corpus for the palestinian arabic dialect," *JLRE*,

[2] K. E. H. et al., "Curras + baladi: Towards a levantine corpus," in *LREC 2022*.

[3] M. M. et al., "Ldc standard arabic morphological analyzer (sama) version 3.1," *LDC2010L01*, Jul. 2010.

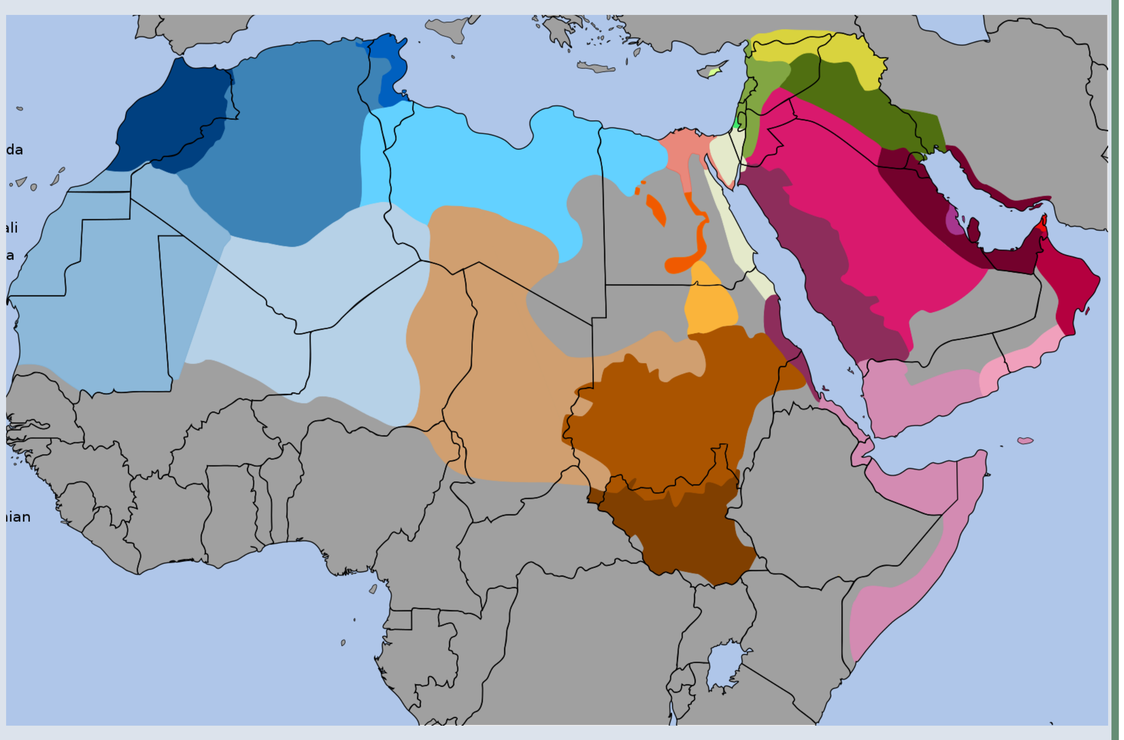
[4] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, 2015.

Arabic is low resourced with several dialects

Modern standard Arabic (MSA):
News, education, media – has more resources

Classical Arabic:
– traditional and literature

Dialects: spoken daily, social media
– region and interlingual influence
Dialect families:
Moroccan, Egyptian, Sudanese, *Levantine*, Iraqi and Khaliji (Gulf)



http://en.wikipedia.org/wiki/Arabic_dialects

CODA annotation guidelines

CODA التهجئة الصحيحة

- The correct and unified spelling of the token
- Followed Curras CODA

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

Prefix annotation guidelines

Prefix(es) سوابق

- Based on SAMA[3] prefix tagset

CONJ { و ، ف }	وحكي، فحكي	PART_RESTRICT { الا }	والاهي
CONNEC_PART { ف }	فعلشان	PROG_PART { ب }	بنطلع
DEM_PRON { ه }	هالبلد	REL_PRON { ما ، ل }	ماشا ، لميين
DET { ال }	الفتية	SUB_CONJ { و ، ما ، إن }	وشكلوا، ماانتي، إنشاء
EMPHATIC_PART { ل ، ت }	لاوريكي، تاوريكي	VOC_PART { يا }	يامعلم
FUT_PART { ح ، س }	حريح، سيريح	IV1S { ا ، آ ، ت }	بامزمز، آخدها ، بتذكر
INTERROG_PART{لا، ما، شو، *أ ، +ما}	مالقاش ، شويدك، اغنيها ، باغنيها	IV2S { ت }	بتدخل
JUS_PART { ل }	ليقول	IV2P { ت }	بتذكروا
NEG_PART { لا ، ما ، مش }	لاشي، مابقتش، مشلازم	IV3S { ت ، ي }	بتبكي، بيضحك
PREP { ل ، ب ، ض ، عـ ، علـ ، *ف ، ت ، ك ، مذ }	لشاحباتها، بالبلد، مشان، عشان، علحصان فالبلد، تاحكيلك، كـكنجي، مذشان	IV3P { ي }	بيجيبوا
		IV1P { ن }	بنطلع

Stem annotation guidelines

<Stem الساق /POS> schema: Used Curras stems
POS, Person, Aspect, Gender, Number

- Same SAMA tagsets

Suffix annotation guidelines

- Suffix(es) لواحق Based on SAMA prefix tagset

PREP { لـ }	صحلي	NSUFF_FEM_PL { ة }	اللبنانية
NEG_PART { ش ، اش ، وش }	بدش، ملقاش، معيوش	PVSUFF_DO:3P {*هن، +هُنّ}	هددهن، هددهُن
REL_ADV {بش}	قاريش	PVSUFF_DO:2P{ *كو، *كِن، +كُن }	عرفتكو، عرفتكِن،عرفتكُن
POSS_PRON_1S { ني }	بعدني	PVSUFF_DO:2FS { كي }	شفتاكي
POSS_PRON_2P { *كو، *كِن، +كُن }	بالكو، بالكِن ، بالكُن	PVSUFF_SUBJ:3P {نو}	طعصفوهن
POSS_PRON_3S { ت }	تاعت	PVSUFF_SUBJ:3S {يت}	طايقيتني
POSS_PRON_3P {*هن، +هُن}	تاعتهن، تاعتهُن	IVSUFF_SUBJ:3S { يت }	معجيبتني
POSS_PRON_2FS { كي }	عتبكي	IVSUFF_DO:2P { *كو، *كِن، +كُن }	بتقلكو، بتقلكِن، تقلكُن
PRON_2FS { كي }	عليكي	IVSUFF_ SUBJ:2P { *كو، *كِن، +كُن }	بديكو، بديكِن، بديكُن
PRON_2P { *كو، *كِن، +كُن }	نيالكو، نيالكِن، نيالكُن		

Curras revisions

- Verified Person, Aspect, Gender, and Number features
- Revised almost all annotations
- Generated a *minimum* set of *unique morphological solutions*
- Used it to annotate Baladi

Tokenization and POS

- No parsing errors
- Buckwalter transliterations
- CODA corresponds to prefixes-stem-suffixes
- Prefix/suffix(es) in the predefined sets
- Stem/POS in the SAMA POS tagset.

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)

Inter-annotation agreement Cohen-*kappa* and F1-score[4]

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

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

Conclusion: A more Levantine Corpus (65.6K tokens)

Inter-annotator agreement on most features illustrates 78.5% Kappa and 90.1% F1-Score.