



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

جامعة نيويورك أبوظبي



NYU ABU DHABI



مختبر كامل
CAMEL Lab

Beyond Cairo: Sa'idi Egyptian Arabic Literary Corpus Construction and Analysis

Mai Mohamed Eida – University of Illinois Urbana-Champaign

Nizar Habash – CAMEL Lab, New York University Abu Dhabi

Egyptian Arabic (EA) Sub-Dialects

- EA is spoken by ~120 million Egyptians
- Ground-truth dialectal surveys classify 4-5 EA sub-dialects (Behnstedt and Woidich, 1985)
- These include Cairene, Da'hlawi, Shar'awi, Sa'idi and differ in phonology, morphology, syntax, and lexicon.
- While EA is the most advanced Arabic Dialect in ArabicNLP, EA resources and applications primarily feature the most prestigious EA sub-dialect, Cairene Egyptian Arabic (CEA).

Egyptian Arabic (EA) Sub-Dialects



Urban Dialects



Rural Dialects

Sa'idi Egyptian Arabic (SEA)

- SEA is spoken by **~25-30 million** Egyptians
- The **2nd most spoken sub-dialect** after CEA
- **Most linguistically distinctive** from CEA (phonology, morphology, syntax, and lexicon)



Rural Dialects

Sa'idi Egyptians

- Sa'idi Egyptians have historically faced **marginalization** for resisting colonial changes in language and religion (Bishai, 1962; Miller, 2003; Nishio, 1994)
- Their dialect is romanticized as a genre in TV shows, but has been **“the most ridiculed, stigmatized, and stereotyped”** otherwise (Bassiouny, 2018).
- **80% live in poverty**, and have the **highest illiteracy rates** in Egypt (World Bank, 2012).
- Al-Said (Upper Egypt) has the highest population of Coptic Christians



Kamel Mahmoud, Guardian of Akhenaten's tomb (Photo credit: Karim Badr)

SEA Corpus



SEA Corpus
Representation



Morphological
Annotation
and Analysis

SEA Corpus Construction

- SEA online users on twitter do not use marked SEA dialectal features online (Eida et al., 2024).
- SEA users avoid positioning themselves as Sa'idi in digital spaces possibly due to the history of social stigma.
- Literature is a non-face threatening environment for language production
- We collect and release a 4 million word corpus of SEA novels and poetry.

SEA Literary Corpus - Poetry

- Two Poets: **Hisham Elgakh** and **AbdelRahman El Abnoudy**
- Both are from **Qena**, a Sa'idi city
- We selected poets who have **published their work in written form** rather than spoken form only
- **355** poems, total of **27k words** across both poets
- We used **OCR to digitize images into text** and manually correct for accuracy



AbdelRahman Al-Abnoudy



Hisham Elgakh

SEA Literary Corpus - Novels

- Self-published literary forum under the “Sa’idi Novels” genre.
- 58 novels by 26 female authors
- 4 authors report a location in Sa’id
- Novels are narrated in Modern Standard Arabic (MSA) and dialogue in DA (SEA and CEA)

روايات وروايات	
قصص رومانسية	
روايات صعيدية	
نوفلا	
قصص مخيفة	
روايات عالمية	
قصص هادفة	
قصص رائعة	
قصص بوليسية	
قصص خيال علمي	
ورشة الكتابة الابداعية	

SEA Corpus



SEA Corpus
Representation



Morphological
Annotation
and Analysis

SEA Representation in SEA Corpus

- **Manual** examination of **16,000 words** from SEA corpus
- We find marked phonological, morphological, and lexical SEA dialectal features **consistent** with the ground truth dialectal surveys (Behnstedt and Woidich, 1985)

	CEA	SEA

SEA Representation in SEA Corpus

- Manual examination of 16,000 words from SEA corpus
- We find marked phonological, morphological, and lexical SEA dialectal features consistent with the ground truth dialectal surveys (Behnstedt and Woidich, 1985)

	CEA	SEA
Phonological	أوي ʔawi very	جوي gawi very

SEA Representation in SEA Corpus

- Manual examination of 16,000 words from SEA corpus
- We find marked phonological, morphological, and lexical SEA dialectal features consistent with the ground truth dialectal surveys (Behnstedt and Woidich, 1985)

	CEA	SEA
Phonological	أوي ʔawi very	جوي gawi very
Morphological	مش هتحرك miʃh+at+ħark not will+move[1.IMPV]	مهتركشي ma+h+at+ħark+ʃi not+will+move[1.IMPV]+not

SEA Representation in SEA Corpus

	CEA	SEA
Phonological	أوي ʔawi very	جوي gawi very
Morphological	مش هتحرك miʃ h+at+ħark not will+move[1.IMPV]	مهتركشي ma+h+at+ħark+ʃi not+will+move[1.IMPV]+not

To conclude this corpus is representative of SEA, **we expect high frequencies of marked SEA features.**

- As in Eida et al (2024), we create a complementary CEA/SEA feature distribution referencing ground-truth dialectal surveys.
- We extract both **CEA and SEA alternations** of the same feature using regexes.

CEA	SEA
أوي ʔawi very	جوي gawi very

- We annotate for correctness in context, and we account for orthographic inconsistencies.
- We **report the prevalence of SEA** features per 10k words.

CEA	SEA
أوي ?awi very	جوي gawi very

75% 25% 100%

- We annotate for correctness in context, and we account for **orthographic inconsistencies**.
- We **report the prevalence of SEA** features per 10k words.

SEA Representation Results

Feature	SEA	CEA	Gloss
Ad1	دلوق، دلوقت	دلوقتي	now
Ad3*	برا	بره، بره	outside
Ad4	برضك، برض	برضه، بردو	also
Ad5	قوي، جوي	اوي	very
Ad6	اهنه، اهنیه	هنا	here
Ad7	اكده، اكديه	كده، كدا	like this
Ad8	لازم، لازماً	لازم	have to
Ad9	بجى	بقى	already
Dem1*	دا	ده	this
Intro2	وين	فين	where
Intro3	ميتى، ميته	امتى	when
Intro4	كيف	ازاي	how
Prep1*	ع	على، ع	on
Prep2*	ف	في، ف	in

SEA Representation Results

Feature	SEA	CEA	Gloss	Tweets	Novel All	Novel Dialogue	NonSEA Authors	SEA Authors	Poetry
Ad1	دلوق، دلوقت	دلوقتي	now	11%	21%	17%	27%	15%	50%
Ad3*	برا	برة، بره	outside	29%	21%	16%	21%	43%	0%
Ad4	برضك، برض	برضه، بردو	also	0%	20%	32%	10%	27%	6%
Ad5	قوي، جوي	اوي	very	0%	59%	56%	58%	26%	0%
Ad6	اهنه، اهنيه	هنا	here	0%	14%	10%	10%	10%	0%
Ad7	اكده، اكديه	كده، كدا	like this	0%	14%	9%	9%	8%	0%
Ad8	لازم، لازماً	لازم	have to	0%	18%	19%	20%	7%	10%
Ad9	بجي	بقي	already	0%	20%	22%	16%	23%	0%
Dem1*	دا	ده	this	24%	37%	28%	53%	73%	0%
Introg2	وين	فين	where	3%	3%	2%	4%	0%	0%
Introg3	ميتي، ميته	امتي	when	0%	33%	24%	31%	21%	100%
Introg4	كيف	ازاي	how	44%	66%	62%	76%	34%	100%
Prep1*	ع	على، ع	on	22%	4%	8%	2%	16%	32%
Prep2*	ف	في، ف	in	24%	6%	12%	2%	13%	23%
Average				11%	24%	23%	24%	23%	23%
Correlation					95.2%				

SEA Corpus



SEA Corpus
Representation



Morphological
Annotation
and Analysis

SEA Morphological Annotation

- How well do **current EA tools** morphologically analyse SEA?
- We present a **semi-automated morphological annotation process for SEA**, following same methodology previously done on Palestinian Arabic (Jarrar et al., 2017) and Emirati Arabic (Khalifa et al., 2018) .

Orthographic Neutralization

- Modern Standard Arabic is the **only variety with a codified, standardized writing system.**

بَرْدُو

b<a>r.du:
also.ADV
also

بَرَضُو

b<a>r.dʕu:
also.ADV
also

بَرَضُهُ

b<a>r.dʕ<u>h
also.ADV
also

CEA

بَرَضُكَ

b<a>r.dʕ<a>k
also.ADV
also

بَرْدُكَ

b<a>r.d<a>k
also.ADV
also

SEA

Orthographic Neutralization

- To address this, we follow CODA* (Habash et al., 2018) guidelines to systematically codify orthographic variation in 15,000 SEA words.
- We emphasize consistency when possible while preserving unique SEA markers
- 2-3 words in every 10 words need modification to align with CODA*, 74% unchanged.
- Modifications include general modifications: replacing ي with ى, and ة with ة but also SEA specific modifications include demonstratives, familial expressions, and adverbials.

Morphological Analysis

- We annotate 4,000 CODA*-annotated SEA words via CAMEL-tools library (Obeid et al., 2020).
- We use CALIMA (morphological analyzer) to generate all possible morphological analyses of each word.
- We also use BERT-Disambiguator in CAMEL-tools to rank analyses based on context.
- We focus on parts of speech, proclitics, enclitics, and gloss.

Evaluation

- SEA is a sub-dialect of EA, and a sister dialect to CEA.

		Accuracy % of Feature								
Total Accuracy of Words		pos	prc0	prc1	prc2	prc3	enc0	enc1	enc2	gloss
81%		88%	100%	99%	100%	100%	98%	100%	100%	84%
OOV Total	16%	53%	100%	96%	100%	100%	96%	100%	99%	51%
INV - Top Choice	81%	100%	100%	100%	100%	100%	100%	100%	100%	100%
INV - Not Top Choice	3%	80%	98%	95%	100%	100%	95%	100%	99%	60%

Conclusion

- We present the first SEA corpus, and validate it for SEA representation.
- SEA literary corpus is moderately representative of SEA compared to current EA resources.
- We find current EA tools perform well for orthographic neutralization and morphological annotation tasks for SEA.

Future Directions:

- Expanding the SEA corpus with additional **spoken and textual content**.
- Develop **automatic tools for processing SEA** in linguistic research as well as Arabic language technologies.
- Developing resources **for underrepresented EA** sub-dialects.
- **Cross-examining spoken SEA** for marked dialectal features compared to written SEA corpus.

Limitations

- SEA is **not a naturally-occurring corpus** and many not accurately represent SEA. Literary works could exaggerate dialectal features.
- The most **comprehensive** dialectal survey was **conducted +30 years ago**.
- While we have surveyed **naturally-occurring data** sources for written SEA text, it is possible there exists **other platforms where users use SEA in written form** that we are not aware of.

Thank you! Questions?
mimm2@illinois.edu