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

Mai Mohamed Eida[1] **and Nizar Habash**[2]
[1]University of Illinois Urbana-Champaign
[2]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
maimm2@illinois.edu, nizar.habash@nyu.edu

Abstract

Egyptian Arabic (EA) NLP resources have mainly focused on Cairene Egyptian Arabic (CEA), leaving sub-dialects like Sa'idi Egyptian Arabic (SEA) underrepresented. This paper introduces the first SEA corpus – an open-source, 4-million-word literary dataset of a dialect spoken by 25 million Egyptians. To validate its representation, we analyze SEA-specific linguistic features from dialectal surveys, confirming a higher prevalence in our corpus compared to existing EA datasets. Our findings offer insights into SEA's orthographic representation in morphology, phonology, and lexicon, incorporating CODA* guidelines for normalization.

1 Introduction

Dialectal Arabic (DA) has been a focus of Arabic NLP throughout the past few decades, the most advanced DA being **Egyptian Arabic (EA)** (Gadalla et al., 1997; Kilany et al., 2002; Maamouri et al., 2014; Jeblee et al., 2014; Fashwan and Alansary, 2021; Habash et al., 2022). EA NLP applications and resources primarily feature the most prestigious EA sub-dialect, **Cairene Egyptian Arabic (CEA),** while sub-dialects such as **Sa'idi Egyptian Arabic (SEA)** are marginalized. As representation within the training data (upstream) influences representation within language technology (downstream), lack of DA sub-dialect resources impacts the representation of DA sub-dialects in Arabic NLP (Dunn, 2020; Tachicart et al., 2022) The focus on CEA over SEA in Arabic NLP is not intentionally biased against SEA, but motivated by the prominence and high accessibility of CEA. SEA speakers tend to avoid using marked dialectal features in online writing (Eida et al., 2024), making it challenging to develop representative textual resources. To address this, we target literature, where SEA speakers intentionally use their dialect, particularly in Sa'idi novels and poetry. This non-face-threatening context allows for the deliberate use of marked features and offers insight into how non-SEA speakers perceive SEA linguistic production.

This paper has three main goals. First, we **collect the first SEA corpus**, a literary dataset of novels, poetry, and short stories, representing a marginalized dialect under-explored in linguistics, literature, digital humanities and NLP. Second, we **assess SEA dialectal feature representation** and find that our corpus better reflects SEA than naturally-occurring tweets. These insights guide efforts to integrate SEA alongside CEA in language technologies and digital humanities research. Finally, we **present a preliminary study on SEA morphological annotation**, a key step toward developing analyzers for NLP and digital humanities tasks that require word-form abstractions due to the morphological richness of Arabic and SEA.[1]

2 Background and Related Work

Modern Standard Arabic (MSA) is the official language of Egypt, and Egyptian Arabic (EA) is the variety spoken among Egyptians. While there is a lot of work on MSA and on EA (in its Cairene variety) (Maamouri and Cieri, 2002; Habash, 2010; Shoufan and Al-Ameri, 2015; Harrat et al., 2017), we focus here on SEA.

2.1 Egyptian Arabic Sub-Dialect Corpora

EA sub-dialects are classified by geographical location, and can be grouped into five sub-dialects (*Cairene*, *Da?hlawi*, *Shar?awi*, *Sa?idi*, and *Badawi*) exhibiting variation across phonol-

[1]We make the texts and our annotations available for research purposes while adhering to copyright guidelines on Github: https://github.com/maimm2/SaidiCorpus2025. The data is mined from public sites, includes only portions of texts, and has scrambled sentence order to address any copyright concerns.

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 292–304

ogy, morphology, syntax, semantics, and lexicon (Behnstedt and Woidich, 1985; Badawi, 1973). CEA and SEA are the most spoken EA sub-dialects, with CEA seen as prestigious and SEA as "the most ridiculed, stigmatized, and stereotyped" (Bassiouney, 2018). Sa'idi Egyptians, comprising 40% of Egypt's population (40 million), have historically faced marginalization for resisting colonial changes in language and religion (Bishai, 1962; Miller, 2003; Nishio, 1994). Despite their numbers, 80% live in poverty, with the highest illiteracy rates in Egypt (World Bank, 2012). Their dialect is often ridiculed, subjecting speakers to discrimination, which discourages them from using SEA online (Eida et al., 2024). This exclusion is reinforced by the lack of language technologies supporting SEA compared to CEA.

There has been limited work on SEA. The most comprehensive linguistic SEA works are ground-truth dialectal surveys by Behnstedt and Woidich (1985) and Khalafallah (1969), two ground-truth surveys from which this paper selects dialectal features to cross-validate representation of SEA in this corpus. As for resources, EA datasets and resources focus on CEA (Gadalla et al., 1997; Kilany et al., 2002; Habash et al., 2012b; Maamouri et al., 2014; Jeblee et al., 2014; Fashwan and Alansary, 2021; Habash et al., 2022). A half-million-word EA corpus and lexicon (Fashwan and Alansary, 2021) reportedly includes SEA data, but it has not been released. Three geo-tagged datasets featuring SEA cities have been published for Arabic sub-dialect identification (Abdul-Mageed et al., 2020a,2020b,2021; Bouamor et al., 2018). However, despite being based on naturally-occurring tweets, these datasets do not adequately capture SEA, as online users may avoid dialectal markers due to historical stigma (Bassiouney, 2014; Bassiouney, 2017). To the best of our knowledge, no existing textual datasets or NLP applications specifically represent SEA.

2.2 SEA Register Variation & Perceptual Dialectology

If CEA users' tweets reflect spoken CEA but SEA users' tweets do not represent spoken SEA, there is a greater distance between the spoken and written registers for Sa'idi Egyptians compared to Cairene Egyptians (Eida et al., 2024). This is further supported by SEA speech in naturally-occurring online videos, which aligns closely with dialectal surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969) to the point where it can be unintelligible to CEA speakers. The absence of marked features in SEA written texts is unexpected and warrants sociolinguistic investigation, highlighting the need for careful validation of EA sub-dialect representation in textual data.

If naturally-occurring written data doesn't reflect SEA, literary texts with clear SEA features provide insight into SEA's written patterns. Assuming SEA digital users deliberately avoid dialectal markers, literature and role-playing offer a non-face-threatening platform for SEA expression. Additionally, Perceptual Dialectology suggests that non-linguists may accurately identify dialect boundaries before linguists (Preston, 1993). While perceptual dialectology focuses on geographic dialect boundaries, examining SEA and non-SEA authors' use of marked SEA features in literary works can inform our understanding of native versus non-native dialect performance (Clark, 2019). This motivates the creation of the first literary SEA corpus presented in this paper, aimed at promoting broader representation of SEA in linguistics, digital humanities and NLP.

3 SEA Literary Corpus Construction

The SEA corpus includes **poetry** and **novels**.

The **poetry** section features works by Sa'idi poets Hisham Algakh and Abdel Rahman el-Abnudi. While more Sa'idi poets exist, many prefer to perform their poetry rather than publish it in books. We selected poets who identify as Sa'idi, perform Sa'idi poetry, and have published their work in textual form, as we are focused on how Sa'idis represent their dialect orthographically. We scan three poetry books from both authors, and use OCR to digitize text from images. We manually correct the OCR digitized text for any erorrs. We plan to include spoken poetry in a future speech corpus.

For the **novels**, we collected works from a self-publishing literary web-forum[2] where authors share their novels across 10 genres such as "Romantic," "Horror," "Sa'idi," "True Crime," and "Science Fiction." Novels are organized by genre and author, and authors may have contributions in multiple genres, and some novels are written as trilogies. Notably, the "Sa'idi" genre is the only culturally specific one, reflecting a trend seen in Egyptian media, where Sa'idi-themed shows and films also exist.

[2] https://stories-blog.com/

	Novels	Poetry
SEA Authors	4	2
Non-SEA Authors	22	0
Documents	58 Novels	355 Poems
Total Words	4,541,835	27,170
Expected SEA Words	1,420,998	12,606

Table 1: SEA Corpus Construction Statistics for SEA located authors, Non-SEA located Authors, total number of documents, total number of words, and approximate number of words extracted from the dialogue of the novels and SEA poems.

We extracted data from the "Sa'idi Novels" subcategory collecting 58 novels by 26 female authors aged 17-40. The dominance of female writers in novel forums is not atypical from other dialectal varieties, such as the Gumar Corpus of Gulf Arabic internet novels (Khalifa et al., 2016). Table 1 summarizes the statistics of our corpus, and a more detailed list appears in Appendix A. Of the 26 authors, only four are located in Sa'idi cities (Asyut, Qena, Sohag, and Southern Egypt), while 22 are based in non-Sa'idi cities or did not report their location. Non-Sa'idi cities include Cairo, Giza, Mansoura, Zagazig, Alexandria, and Damietta. While we refer to authors as SEA authors and Non-SEA authors based on their reported geographical location, but we do not make claims about their identity as Sa'idi or non-Sa'idi. Some novels use MSA for narration and SEA for dialogue (Appendix B Figure 2), while others alternate between CEA for narration and SEA for dialogue (Appendix B Figure 3).

Each novel title follows the same template, making it consistent across the site. This template is "Novel Title" followed by the number of novels in trilogy "Part X" and finally its reference to the author "by Author Y" – "Novel Title Part X by Author Y". The first page of each novel includes a descriptive picture with character(s) and the novel title, introduction or sample of the novel, followed by all the linkable chapters. For the introduction, the author includes approximately 1000 words to introduce the synopsis, characters, and settings. On occasions, this section might contain editor notes on novel organization, spelling, or grammatical errors. If the author does not include an introduction, they include a 500 word sample extracted from the novel as a teaser to the novel. For chapter links, every link to each chapter follows a template of "Novel Title Part X by Author Y Chapter Z". Chapters vary from 1 chapter under the "novella" genre to 56 chapters with an average of 25 chapters per novel. This organizational structure is uniform across all genres, authors, and novels.

The stylistic choices of each are mostly consistent by author. For example, if an author uses MSA to narrate the novel, they are consistent with using MSA in all the novels they write. With 2-3 exceptions, where they use Dialectal Arabic to narrate the novel once, but MSA otherwise. Another example is a punctuation signifier for characters beginning their dialogue, where authors are mostly consistent with either ":" or ": -" or a new line. This is is illustrated in Table 1 in Appendix A.

We release the corpus organized author by author, and novel by novel. We extract the dialogue only for each chapter using dialogue markers in the novel, and we exclude novels where there is no distinction between dialogue and narration as an attempt to isolate the SEA dialect as much as possible from the MSA and CEA used within the same novel. With this, we achieve our first goal of developing and releasing the first SEA corpus. Next, we need to understand how representative it is compared to naturally-occurring data as well as examine the written patterns of SEA that can further guide SEA speech annotation, morphological analysis, and more.

4 SEA Linguistic Features

To explore SEA written production within the corpus, we begin by manual examination of the marked dialectal features of SEA presented in the dialogue of the novels and SEA poems. We examine a randomized sample of 16,000 words across poetry and prose, while cross-referencing marked dialectal features found in the corpus with the ground-truth dialectal surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969). Results show marked SEA features consistent with the ground-truth surveys. Table 2 highlights a sample of the marked phonological and morphological SEA dialectal features found in the corpus consistent with the ground-truth dialectal surveys. We provide a CEA/SEA minimal pair of each feature for comparison, along with IPA transcriptions, transliteration,[3] and an example which includes the text as found in the corpus, transliteration, IPA transcription and gloss. Additionally, we make three general observations on the nature of the literary novels that might

[3] Arabic transliteration is described in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

		CEA Feature		SEA Feature		Example		
		IPA	***Letter***	***IPA***	***Letter***	***CEA***	***SEA***	**SEA IPA / Gloss**
Phonology	**Consonant**	g	ج	d	د d	جمل jml	دمل dml	damal camel
				[d]ʒ	ج	جوزك jwzk	جوزك jwzk	ʒu:z+ik husband + your[2.F.SG]
					ج	جرا grA	چرا jrA	ʒara happen[PFV.3]
		ʔ	ا/ق A/q	g	ج/ق j/q	قوي/اوي Awy/qwy	قوي/جوي jwy/qwy	gawi very
	Vowel Lengthening	i/a	-	i:/a:	ي	كدة kdħ	اكديه Akdyh	ʔikdih like this
	Vowel Shortening	a:/i:	ي	a/i	-	مراتي mrAty	مرتي mrty	marat+i wife+my
		IPA	***Morpheme***	***IPA***	***Morpheme***	***CEA***	***SEA***	**SEA IPA / Gloss**
Morphology	**Future Prefix**	h, ħ	ه،ح-	ʕ,ħ	ع،ح-	هتسكن htskn	عتسكن ʕtskn	ʕa+t+skun will+you[2.SG.MASC]+live[IMPFV]
	Negation	ma-ʃ	م-ش	ma-ʃ	م-شي	مش هتحرك mš htHrk	مهتحركشي mhtHrkšy	ma+h+at+ħark+ʃi not+will+move[1.IMPFV]+not
				-ʃ	م-ش	مجاش mgAš	مجاش mgAš	ma+ʒa:+ʃ not+come[3.SG.MASC]+not
				-ʃi	شي	مروحتولهاش mrwHtwlhAš	روحتولهاشي rwHtwlhAšy	ruħt+tu+l+ha:+ʃi go[3.IMPFV]+you[PL]+to+her+not
	1.P.S. Prefix-Suffix	ʔ	-أ	n-u:	ن-وا	أروح ArwH	نروحوا nrwHwA	n+ru:ħ+u: [1.SG]+go[1.SG.IMPFV]

Table 2: Sample of Phonological and Morphological SEA Marked Dialectal Features as observed in the SEA corpus compared to their CEA counterparts in line with ground-truth dialectal surveys.

affect SEA representation.

First, we find that some authors alternate between SEA and CEA variations of some features within the same novel. For example, SEA authors alternate between CEA feature كدا *kdA* /kida/ and SEA feature إكده *Ăkdh* /ik.dih/ meaning 'like this'. This could indicate masking of one feature and substituting with the other, or could be explained by our next observation, where characters are assigned different dialects within the context of the novel. Second, an interesting theme across SEA authors is assigning more marked SEA features to the speech of elders, and less marked SEA dialectal features to speech of young characters. This could explain the lack of marked SEA features in naturally-occurring data in digital settings, given the younger demographic use social media platforms more frequently (Kindt and Kebede, 2017). This also poses the question: *are Sa'idi Egyptian youth moving away from SEA marked dialectal features compared to older generations?* This would require further research. Third, Non-SEA authors orthographically exaggerate SEA dialectal features to a larger extent than SEA authors. For example, CEA authors use چ *ž* to represent the SEA sound ʒ. This is a marked Persian letter when used in Arabic, since it is not part of the MSA script. SEA authors use the MSA ج *j* to represent the same sound. Both SEA and some non-SEA authors use ج *j* to represent SEA g, such بجي bjý /baga/ 'already' instead of بقى bqý /baga/ 'already', however, the frequency of بجي bjý /baga/ 'already' is much higher in non-SEA authored novels. While SEA authors are moving away from using marked dialectal features, non-SEA authors usage and perception of SEA marked dialectal features confirm their alignment with the ground-truth dialectal surveys.

Along with the table above, we also observe common differences in some verb patterns, specifically CEA verb pattern V 'itCVCXVC' (e.g. اتكلم *Ătklm* /itkallim/ 'he spoke' and اتجوز *Ătjwz* /itgawwiz/ 'he got married'). In both cases the 't' in the pattern is assimilated to produce اكّلم *Ăklm* /ikkallam/, and اجّوز *Ăjwz* /iʒʒawwaz/, respectively.

In addition to the discussed SEA features, we find that the SEA corpus carries a high number of lexical items unique to SEA, with some MSA and Coptic etymology. This is consistent with the literature, which indicates that Upper Egypt did not fully transition to Arabic until the 17th century (Bishai, 1962; Lipiński, 1997; Soliman, 2007), 7 centuries after the Delta and Cairo area did, and therefore SEA retains heavier influence of both MSA and Coptic in lexicon. Words with MSA origin include حديت *Hdyt* /ħadi:t/ 'speech/conversation', and زين *zyn* /zi:n/ 'good', and words with Coptic origin include عفشة *ʕfšħ* /ʔifʃa/, and شينة *šynħ* /ʃe:na/ both meaning 'bad/ugly'. This qualitative analysis sheds insight into answering our next question: *is this corpus quantitatively representative of SEA?*

5 How Well does SEA Literary Corpus Represent SEA Dialects?

5.1 Methodology

If this corpus is representative of SEA, we expect high frequencies of marked SEA dialectal features rather than marked CEA dialectal features. To measure prevalence of SEA dialectal features in this corpus, we adopt the 'SEA Ground-Truth Dialect Features' methodology in Eida et al. (2024). These features, which include morphological and lexical features of each sub-dialect, are selected from ground-truth dialect surveys and used as a distance measure between spoken and written SEA. Features include demonstratives, interrogatives, prepositions, and adverbs, and as reported in SEA dialectal surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969; Leddy-Cecere and Schroepfer, 2019). Our motivation is to select features where there is a distinction between SEA and other EA sub-dialects in orthography, yet are essential to the syntax of SEA.

We create a complementary CEA/SEA feature distribution where we extract both CEA and SEA alternations of the same feature, and report the prevalence of SEA features. For example, if we search for the alternation of the adverb 'now', the CEA alternation would be دلوقتي *dlwqty* /dilwaʔti/ and the SEA alternation would be دلوق *dlwq* /dilwagˁ/ or دلوقت *dlwqt* /dilwagt/. Using regexes, we string match both orthographic representations, and manually annotate in context for correctness. After removing incorrect matches, we measure the frequency of each alternation per 10k words. Since the features are complementary, if the SEA feature is reported as 25% in the results shown in Table 3, the remaining 75% would be the CEA alternation of the same feature. This would indicate that the CEA alternation is more prevalent in the corpus than the SEA alternation.

We modify the the 'SEA Ground-Truth Dialect Features' adopted from Eida et al. (2024) to reflect the qualitative results established in section 4. We confirm the presence of features used in Table 3 in the corpus, and account for their varying orthographic representations in SEA data.

After extracting the remaining features adopted from Eida et al. (2024), and checking for false positives, we remove any features that result in a 0% across all corpora despite their existence in the ground-truth dialectal surveys. Otherwise, all possible orthographic variations are accounted for in feature extraction, such as interchangeably using ي *y* ى *ý* and ه *h* ة *ħ* since these substitutions are common in written Dialectal Arabic. While the selected features have limitations in detecting SEA dialect markedness, they provide insight into SEA representations across key features, as shown by the results.

For consistency, we compare against the Micro-Dialect, NADI2020, NADI2021 (Abdul-Mageed et al., 2020a, 2020b, 2021) SEA cities' datasets following Eida et al. (2024)'s methodology, with a tweet corpus of 73,404 words. This dataset has been reported to be non-representative of SEA features (Eida et al., 2024), and would be a good baseline to compare against the SEA representation of this corpus, especially as we modified some marked dialectal features. *Is this corpus more representative of SEA than naturally-occurring tweets?* To answer this, we compare SEA dialectal feature usage and prevalence across all novels, novel dialogue only, SEA authored novels, non-SEA authored novels, and poetry, as illustrated in Table 3. Results are reported with a focus on SEA feature alternations.

5.2 Results

Consistent with the findings of Eida et al. (2024), SEA features are less prevalent in the Tweet corpus compared to the SEA literary corpus. The most marked SEA dialectal features added to 'SEA Ground-Truth Dialect Features' after qualitative analysis seem to be non-existent in the Tweet corpus, with a consistent 0% across Ad4-Ad9. While

				SEA Feature Prevlance					
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	لازمن، لازماً	لازم	have to	0%	18%	19%	20%	7%	10%
Ad9	بجى	بقى	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	ميتى، ميته	امتى	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	30.6%	95.2%		59.5%		

Table 3: Share of SEA Dialectal Features in SEA Tweet Corpus, SEA literary Corpus, Novel Dialogue, and in SEA vs. non-SEA Authored novels, and Poetry. * indicates the least SEA marked dialectal features.

this also seems to be true for Poetry, Poetry features 100% for two of the most marked SEA dialectal features at Introg3 & Introg4. Despite both Tweets & Poetry's sample size, SEA marked dialectal features is more prevalent in Poetry than Tweets. Since both poets identify as from Qena, the results show agreement in the marked SEA dialectal features they use. At first glance, we can conclude the SEA literary corpus presented is moderately more representative of SEA with an average of 24%, while Tweet corpus average of 11% is not, with the exception of Prep1 and Prep2.

One explanation of higher frequencies of SEA alternations of Prep1 and Prep2 could be that the Tweet corpus is naturally-occurring, therefore users do not adhere to MSA writing standards expected in writing literary texts such as novels. The standard orthographic representation في *fy* /fi:/ 'in' is used more frequently in the literary corpus than Tweet corpus, while SEA Tweets show ف *f* /f/ 'in' more frequently. It could also be that because Prep1 and Prep2 are the least marked SEA features in this table, Non-SEA authors are not aware of its subtle SEA markedness caused by removing the final letter in the preposition. In support of this prediction, Table 3 shows SEA authors also use the predicted Prep1's SEA alternation represented as ف *f* /f/ 'in' with a report of 16%, compared to usage among non-SEA authors at only 2%. This strongly suggests ف *f* /f/ 'in' is the orthographic representation preferred in SEA, consistent with the reported ground-truth dialectal surveys, despite it being the least marked SEA feature in this list of marked SEA features.

The most prevalent SEA dialectal features in the SEA literary corpus shown in Table 3 is Ad5 قوي ، جوي *qwy, gwy* /gawi/, Introg4 كيف *kyf* /ke:f/, Introg3 ميته ، ميتى *mytý* , myth /me:ta/ and Dem1 دا daA /da:/. This is in line with the reported features of the ground-truth dialectal surveys, however, SEA Introg2 وين *wyn* /we:n/ 'where' seems to be almost non-existent across all corpora except for Poetry, despite being reported in the ground-truth dialectal surveys. SEA usage appears to be shifting toward the CEA فين *fyn* /fe:n/ 'where', as suggested by the prevalence of the CEA Introg2 alternation, which is captured in over 97%+ of the extracted cases in this corpus. However, this should be confirmed with naturally-occurring, more representative spoken SEA corpora. On the other hand, the results for SEA and Non-SEA authors in Ta-

Phenomenon	Text	CODA*	Gloss
Negation Clitics	مجاش mjAš	ما جاش mA jAš	neg come[3.IMPFV]+neg
Prepositional Clitics	قالولي qAlwly	قالوا لي qAlwA ly	tell[3.PL.PFV] me
Familial Expressions	اما، امه، امى AmA, Amh, Amý	اما AmA	mother
	بت bt	بت bt	daughter
Ta Marbuta	مرت عمي mrt ςmy	مرة عمي mrħ ςmy	uncle's wife
Relative Pronouns	ال Al	اللي Ally	that
Existentials	في fy	فيه fyh	there is
Demonstratives	اها AhA	اها AhA	this
Adverbials	اهني، اهنه Ahny, Ahnh	اهنه Ahnh	here
	اكده، اكديه، اكدي Akdh, Akdyh, Akdy	اكده Akdh	like this

Table 4: Most Common Modification in SEA corpus to follow CODA* guidelines

ble 3 are mixed. SEA alternations of Pron1, Pron2, Ad1, Introg3, Dem1, and Ad5 show prevalence in SEA authors' novels more then non-SEA authors. The orthographic representation of Dem1 دا *daA* /da:/ seems to be accurately representative of ground-truth dialectal surveys reports of ending in long vowel /a:/ as opposed to the CEA ending Dem1 ده *dh* /dah/. The mixed results are expected, as both SEA and non-SEA authors are writing novels in SEA. We conclude that both are moderately representative of SEA, more than existing EA geotagged datasets.

One final question remains: are authors consistently writing in SEA? Could it be the larger majority of authors are impacting the reported SEA feature prevalence? The correlation between the prevalence of SEA marked features in Novel All and Novel Dialogue is high as expected, but the correlation between SEA feature prevalence across Non-SEA and SEA authors is relatively lower. There are differences in the consistency and choices made by SEA and non-SEA authors in representing SEA marked dialectal features, and we visualize the distances between SEA and non-SEA authors specific corpora SEA usage in Figure 4 included in Appendix C. In other words, the disconnect between SEA corpora and expected SEA features might be a result of individual differences across author writing styles, with non-SEA authors aligning closely to a specific SEA usage compared to SEA authors. This leads to the conclusion that there is a SEA representation distinction depending on location, with a gap between SEA and non-SEA author usage of marked SEA dialectal features. In conclusion, the SEA literary corpus exhibits higher frequencies of marked SEA features compared to the baseline Twitter corpus. This is consistent with the ground-truth dialectal surveys.

6 Towards a Morphologically Annotated SEA Corpus

In this section, we present a preliminary study partially automating SEA morphological annotation using existing EA morphological analysis tools to streamline SEA morphological analysis and annotation. Following the methodology in Khalifa et al. (2016) and Jarrar et al. (2017) by using CODA* (Habash et al., 2018) and CALIMA EGY (Habash et al., 2012b), we present a semi-automated morphological annotation process for SEA, with expected modifications and results.

6.1 Orthographic Neutralization

Modern Standard Arabic (MSA) is the only Arabic variety with a standardized codified writing system (Brustad, 2017; Håland, 2017; Høigilt and Mejdell, 2017). For Dialectal Arabic (DA) generally, there is no standardized orthographic system, which presents one of the main challenges in the Arabic NLP. Written EA output is orthographically inconsistent, across the same lexical items due to the complex nature of Arabic orthography and the intertwined nature of Arabic vowel diacriticization rules, standardized MSA, and DA marked orthographic representations for features exclusive to DA result in the complexity of parsing DA orthog-

		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%

Table 5: Accuracy of Morphological Analysis and Tagging of SEA Data based on Total, Out of Vocabulary (OOV) words, and In Vocabulary (INV) words.

raphy. To address the DA orthographic inconsistencies and its effect on DA parsing, there have been several NLP DA codification guidelines, including CODA and CODA* (Habash et al., 2012a; Habash et al., 2018) DA guidelines aim at systematically codifying DA orthographic variation, emphasizing consistency when possible to facilitate DA parsing, while preserving the unique DA markers for each dialect. For the EA dialect, CODA* primarily accounts for sub-dialects spoken in Cairo, Alexandria, and Aswan. In this paper, we add SEA to the CODA* DA map. First, we annotate and release 15,000 SEA words from the corpus using CODA* to be used as a reference along with CODA* rules to codify SEA data. Our results show 2-3 in every 10 words need modification to align with CODA* rules, with 74.7% words unmodified. This falls within the comparable range of CODA* annotation results for Palestinian Arabic (Jarrar et al., 2017) at 86.54%, as well as Emarati Arabic (Khalifa et al., 2018) at 78.1%. Aside from SEA marked lexical items, the most common modifications are listed in Table 4. Other modification heavily featured is substituting letters such ة، ه and ي، ى with one another based on morphophonetic and morphosyntactics of the word. For example, ة is a suffix which denotes the feminine gender for nouns, and a noun such as حاجه *HAgh* /ħa:ʒa/ meaning 'thing', must be written as حاجة *HAjħ* /ħa:ʒa/ also meaning thing, but the ة ħ is consistent with following CODA* rules in indicating the gender of the noun and in accordance with how this would be written in MSA as well.

6.2 Morphological Analysis

We further annotate 4,000 CODA*-annotated SEA sentences using CALIMA's morphological analyzer (Habash et al., 2012b) and BERT-Disambiguator (Inoue et al., 2022) via CamelTools (Obeid et al., 2020). CALIMA's analyzer generates all possible morphological interpretations for each sentence, while the BERT-Disambiguator ranks these interpretations based on context. We then select and annotate parts-of-speech, proclitics, enclitics, and English glosses from the output of both tools.

6.3 Evaluation

We conduct an evaluation on the quality of CALIMA's automatic morphological analysis on SEA data. Given that SEA is a sister dialect to CEA, a dialect that CALIMA models, we predict the performance on SEA will be relatively high due to the overlap between both dialects as well as CODA* disambiguating some of the orthographic representations in the SEA data. As illustrated in Table 5, we check for accuracy of POS, proclitics, enclitics, and gloss. We measure "Total Accuracy" by accuracy of all features. We measure "OOV Total" for the remaining features if 1 or both POS and Gloss features are OOV. We measure "INV - Not Top Choice" if both POS and Gloss features are found within the list of generated outputs of the morphological analyzer, but not selected by the BERT-Disambiguator as the top choice in context.

The overall accuracy for SEA data is at 81% and is promising given the lack of current morphological analysis tools trained on SEA data. The remaining 19% contain 16% OOV words, where the largest error rate was observed in English gloss. This is expected: SEA lexical items retain MSA etymology and overlap with CEA morphological features, yet denote different semantic representations. For example, علامها ʕlAmhA /ʔala:mha/ 'her education' is correctly in POS as noun, segmented as علام + ها identifying the clitic as '3fs_poss', yet incorrectly glossed as 'expert'. It is worth noting that this gloss is not too far fetched given they share the same Arabic root ع ل م. Other error analysis

reflects the qualitative results in Table 2. One example is no analysis is generated for verbs with the future clitic ع ح /ʔ/ 'will.[FUT]', as the future clitics in CEA are ه *h* /h/ and ح *H* /ħ/ only. Our evaluation results are comparable to those of Palestinian Arabic (Jarrar et al., 2017) and Emarati Arabic (Khalifa et al., 2018), both of which also use EGY morphological analyzers to semi-automate their respective DA corpora.

7 Conclusions and Future Work

This paper presents the first SEA corpus, including its construction and analysis of SEA representation. We find the corpus moderately representative of SEA, though its consistency across authors is influenced by variation within SEA sub-dialects (Behnstedt and Woidich, 1985). Despite this, the corpus offers valuable insight into the phonological, orthographic, lexical, morphological, and variations between SEA and CEA.

Future work will focus on expanding the SEA corpus with additional spoken and textual content, as well as manual annotations to improve its consistency, representation, and overall usability. We also plan to develop automatic tools for processing SEA to support broader linguistic research and application development.

Limitations

This corpus, being literary and not naturally-occurring, may not accurately represent SEA writing practices, as literary works often exaggerate dialectal features. Additionally, since the ground-truth dialectal surveys are over 30 years old, some language changes may have occurred, making certain features less representative of current SEA dialect. We have surveyed other naturally-occurring data sources to validate the presence of SEA features, however, there seems to be very limited instances where online users produce written SEA. We have found most SEA production is in speech, and delivered via video, however, it is possible there exist other platforms where users produce written SEA that we are not aware of.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

El-Said Badawi. 1973. *Mustawayat al-Arabiyah al-muasirah fī Misr : bahth fi alaqat al-lughah bi-al-hadarah*. Dār al-Mārif, Cairo.

Reem Bassiouney. 2014. *Language and identity in modern Egypt*. Edinburgh University Press.

Reem Bassiouney. 2017. *Identity and dialect performance: A study of communities and dialects*. Routledge.

Reem Bassiouney. 2018. Constructing the stereotype: Indexes and performance of a stigmatised local dialect in Egypt. *Multilingua*, 37(3):225–253.

Peter Behnstedt and Manfred Woidich. 1985. Die ägyptisch-Arabischen dialekte. *Tübinger Atlas des Vorderen Orients/Beihefte/B*, 50.

Wilson B Bishai. 1962. Coptic grammatical influence on Egyptian Arabic. *Journal of the American Oriental Society*, 82(3):285–289.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kristen Brustad. 2017. Diglossia as ideology. In *The Politics of Written Language in the Arab World*, pages 41–67. Brill.

Urszula Clark. 2019. *Staging language: Place and identity in the enactment, performance and representation of regional dialects*, pages 1–22. De Gruyter Mouton, Berlin, Boston.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Mai M. Eida, Mayar Nassar, and Jonathan Dunn. 2024. How well do tweets represent sub-dialects of Egyptian Arabic? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects*.

Amany Fashwan and Sameh Alansary. 2021. A morphologically annotated corpus and a morphological analyzer for egyptian Arabic. *Procedia Computer Science*, 189:203–210.

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.

Nizar Habash, Mona Diab, and Owen Rambow. 2012a. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 711–718, Istanbul, Turkey.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012b. 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, Reham Marzouk, Christian Khairallah, and Salam Khalifa. 2022. Morphotactic modeling in an open-source multi-dialectal Arabic morphological analyzer and generator. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 92–102.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Eva Marie Håland. 2017. Adab sākhir (satirical literature) and the use of egyptian vernacular. In *The politics of written language in the Arab world*, pages 142–165. Brill.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2017. Machine translation for Arabic dialects (survey). *Information Processing & Management*.

Jacob Høigilt and Gunvor Mejdell. 2017. *The politics of written language in the Arab world: Writing change*. Brill.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51:745–775.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 196–206, Doha, Qatar.

Abdelghany A Khalafallah. 1969. *A descriptive grammar of Saidi Egyptian colloquial Arabic*, volume 32. Walter de Gruyter GmbH & Co KG.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Kristian Takvam Kindt and Tewodros Aragie Kebede. 2017. A language for the people?: Quantitative indicators of written darija and ammiyya in Cairo and Rabat. In *The Politics of Written Language in the Arab World*, pages 18–40. Brill.

Thomas Leddy-Cecere and Jason Schroepfer. 2019. *Egyptian Arabic*, pages 433–457. Routledge.

Edward Lipiński. 1997. *Semitic languages: outline of a comparative grammar*. Peeters Publishers.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 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-2014)*. European Language Resources Association (ELRA).

Mohamed Maamouri and Christopher Cieri. 2002. Resources for Natural Language Processing at the Linguistic Data Consortium. In *Proceedings of the International Symposium on Processing of Arabic*, pages 125–146, Manouba, Tunisia.

Catherine Miller. 2003. Variation and change in Arabic urban vernaculars. In *Approaches to Arabic Dialects*, pages 177–206. Brill.

Tetsuo Nishio. 1994. *The Arabic dialect of Qift (Upper Egypt): grammar and classified vocabulary*, volume 27 of *Asian & African Lexicon*. Institution for the Study of Languages and Cultures of Asia and Africa.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for

Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032.

Dennis R Preston. 1993. Folk dialectology. *American dialect research*, pages 333–378.

Abdulhadi Shoufan and Sumaya Al-Ameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, page 36, Beijing, China.

Mary Soliman. 2007. *Arabic Dialectology and the Influence of Coptic on Egyptian Arabic*. Ph.D. thesis, Florida Atlantic University.

Ridouane Tachicart, Karim Bouzoubaa, Salima Harrat, and Kamel Smaïli. 2022. Morphological Analyzers of Arabic Dialects: A survey. *Studies in Computational Intelligence*, 1061.

World Bank. 2012. Young People in Upper Egypt.

A Detailed Corpus Data

Author #	Demographic	Novels #	Chapters	Narration	Dialogue Marker
1	Unknown	1	53	MSA	:
		2	41	MSA	:
		3	34	MSA	N/A
2	SEA	1	36	DA	:
		2	23	DA	:
		3	31	DA	:
3	Non-SEA	1	47	MSA	:
4	SEA	1	42	DA	-/:
		2	56	DA	-/:
5	Non-SEA	1	40	MSA	:
6	Non-SEA	1	40	MSA	: - / :- / -
		2	9	MSA	: - / :- / -
7	Non-SEA	1	20	MSA	:
		2	22	MSA/DA	:
		3	26	MSA/DA	:
		4	25	MSA/DA	:
8	Unknown	1	45	DA	:/: -/\n
		2	30	DA	:/: -/\n
9	Non-SEA	1	40	MSA	•/.
		2	40	MSA	.../..
10	Unknown	1	33	MSA	-/:
11	Unknown	1	39	MSA	:
		2	30	MSA	:
12	Unknown	1	20	MSA	:\n
		2	30	MSA	:\n
		3	35	MSA	:\n
13	Non-SEA	1	41	MSA	:
14	Unknown	1	36	MSA	:
15	Non-SEA	1	30	MSA	:/\n/:\n
		2	31	MSA	:\n-
16	Non-SEA	1	20	MSA	:
17	Non-SEA	1	28	MSA	:
		2	21	MSA	:
18	Unknown	1	20	MSA	=
19	Non-SEA	1	24	MSA	-/: -
		2	7	MSA	-/: -
		3	20	MSA	:
20	Unknown	1	16	MSA	: -
		2	16	MSA	: - /\n
		3	21	MSA	-/\n/
		4	20	MSA	: - /\n
		5	41	MSA	: - /\n
		6	39	MSA	:
		7	27	MSA	:-
21	Non-SEA	1	30	MSA	\n
22	Non-SEA	1	22	MSA	: \n
		2	20	MSA	\n/?/:
23	Unknown	1	30	MSA	:
		2	20	MSA	:
24	SEA	1	27	MSA	:
		2	20	MSA	:
		3	20	MSA	:
25	Unknown	1	7	MSA	:
26	SEA	1	20	MSA	""/-/_/{}
		2	19	MSA	:/\n
		3	26	MSA	:/\n
		4	24	MSA	:/\n

Figure 1: Detailed corpus data organized by demographic, author number followed by each novel they wrote and the number of chapters in each novel. We also add the dialect used for narration either as Modern Standard Arabic (MSA) or Dialectal Arabic (DA), and the dialogue markers each author used to seperate narration from the narration in each novel.

B Text Samples

استوقفته شهيرة التي لم ولن تتغير قائلة: وبعدهالك ياولد أبوك،
مهتسمعش كلامي وتشوف العروسة..

Figure 2: Sample of using MSA for narration (underlined) and SEA for Dialogue from SEA Corpus. Translated as *"Shahira, who has and will never change, stopped him saying: What now, child? Are you still not going to listen and meet the bride.."*

الباب خبط وكانت خديجه دخلت وهي باصه في الارض بخوف و
ماسكه تيشيرت في ايديها سليمان: خير يا خديجه يا بنتي.

Figure 3: Sample of using CEA for narration (underlined) and SEA for Dialogue from SEA Corpus. Translated as *"After a knock on the door Khadeeja entered while looking at the floor in fear. She held the t-shirt in her hand. Siliman: what is it, Khadeeja, my daughter."*

C Principal Component Analysis of SEA and Non-SEA Feature Usage by Author

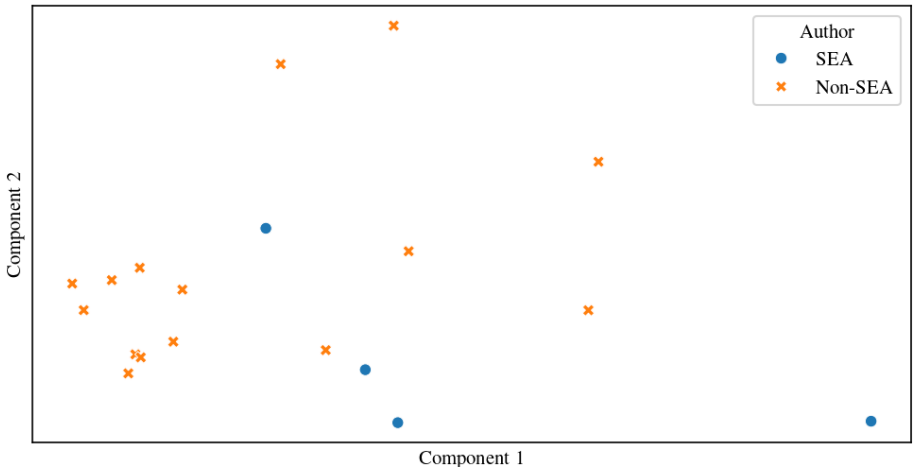

Figure 4: Author-by-author plots of SEA and Non-SEA feature usage, visualized using PCA for dimension reduction. The original vectors undergoing PCA are the relative frequency of SEA and Non-SEA dialectal features. It is clear that authors taken to represent both SEA (circles) and Non-SEA (x's) are not intermingled. This would indicate SEA and Non-SEA feature prevalence seems to be because some authors emphasize SEA selected features more than others. Non-SEA author usage seems to be organized around a consistent representation of SEA, indicate by the cluster of x's to the left.