How Well Do Tweets Represent Sub-Dialects of Egyptian Arabic?

Mai Mohamed Eida[1], Mayar Nassar[2], Jonathan Dunn[1],
[1]University of Illinois Urbana-Champaign, [2]Ain Shams University
Correspondence: maimm2@illinois.edu

Abstract

How well does naturally-occurring digital text, such as tweets, represent sub-dialects of Egyptian Arabic (EA)? This paper focuses on two EA sub-dialects: Cairene Egyptian Arabic (CEA) and Sa'idi Egyptian Arabic (SEA). We use morphological markers from ground-truth dialect surveys as a distance measure across four geo-referenced datasets. Results show that CEA markers are prevalent as expected in CEA geo-referenced tweets, while SEA markers are limited across SEA geo-referenced tweets. SEA tweets instead show a prevalence of CEA markers and higher usage of Modern Standard Arabic. We conclude that corpora intended to represent sub-dialects of EA do not accurately represent sub-dialects outside of the Cairene variety. This finding calls into question the validity of relying on geo-referenced tweets alone to represent dialectal differences.

1 Egyptian Arabic Sub-Dialects

Existing work on Egyptian Arabic (EA) sub-dialects primarily uses geo-referenced data to represent specific varieties. The question here is whether existing EA corpora adequately represent the intended sub-dialects: do existing written corpora of EA equally represent both majority varieties (e.g., Cairene Egyptian Arabic: CEA) and minority varieties (e.g., Sa'idi Egyptian Arabic: SEA)? This is an important question for two reasons: first, representation within the training data (upstream) influences representation within language technology (downstream). This means that dialect adaptation for less prestigious varieties like SEA depends on these dialects being adequately represented in training corpora (Biber, 1993; Dunn, 2020). Second, spoken and written register variation in Arabic can impact dialect representation. For example, results in this paper suggest that speakers of CEA freely use their dialect in tweets but speakers of SEA revert to Modern Standard Arabic (MSA). This implies that the relationship between dialect and register is not predictable across sub-dialects.

Current work on Dialectal Arabic (DA) resources and applications does not take into account DA variation beyond the country level (Abdul-Mageed et al., 2020a, 2020b, 2021; Bouamor et al., 2018; Tachicart et al., 2022). Further, this work has not considered spoken and written register variation across sub-dialects. Therefore, this paper addresses two specific questions. First, which Egyptian Arabic sub-dialects are represented within existing digital written corpora, specifically tweets? To find out, we compare these corpora with ground-truth dialect surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969). Second, is EA in a digital written register, specifically tweets, equally representative of spoken sub-dialects? To find out, we compare the relative usage of DA vs. MSA features in tweets across two sub-dialects of Egyptian Arabic.

If current datasets are representative of EA sub-dialects, and if register variation across written and spoken EA is limited, then NLP tasks like Arabic micro-dialect identification, machine-translation, and morphological parsing can be adapted for dialectal varieties using Tweet-based corpora. In other words, this would mean that digital written data, as a register, remains representative of EA sub-dialects. However, if the current datasets are not representative of EA sub-dialects, this means that sub-dialects (beyond CEA) are low-resource and that more data collection is needed to represent all EA sub-dialects. Further, the possibility that digital written registers are not equally valid for all sub-dialects means that other sources of EA sub-dialect data, such as speech, should be explored. In other words, if speakers of less prestigious dialects like SEA revert to standardized forms in written registers then spoken data must also be used for dialect adaptation.

The primary contribution of this paper is to measure how well written registers represent different

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 41–55
June 20, 2024 ©2024 Association for Computational Linguistics

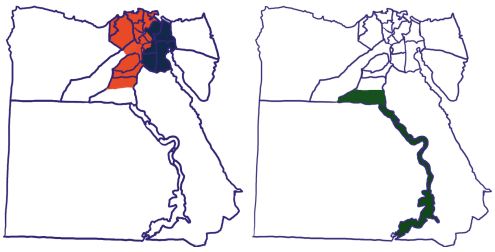

Figure 1: Map of Egyptian Arabic Sub-dialects (Woidich, 1996): CEA and other rural dialects (left), SEA dialect (right)

sub-dialects of Egyptian Arabic as well as the validity of specific datasets designed to capture these sub-dialects. We find that the more prestigious CEA is well-represented but that resources intending to represent the less prestigious SEA fail to do so.

The outline of this paper is as follows: Section 2 provides an overview of Egypt's sub-dialects, addresses register variation in Arabic, and previous work on sub-dialect data collection and dialect identification. In Section 3, we provide an overview of specific features of two Egyptian sub-dialects: Cairene Egyptian Arabic (CEA) and Sa'idi Egyptian Arabic (SEA). These features are drawn from ground-truth dialect surveys. Section 4 discusses a baseline corpus, a large reference corpus to which other datasets are compared, and three sub-dialect corpora. The remaining sections present results and discuss the significance of representation and register variation across CEA and SEA.

2 Related Work

2.1 Egyptian Arabic Sub-dialects

Arabic is a diglossic language (Ferguson, 1959), with Modern Standard Arabic (MSA) considered the High-variety, and Dialectal Arabic (DA) the Low-variety. While MSA is the official language of Egypt, Egyptian Arabic (EA) is the dialectal variety spoken among Egyptians. EA sub-dialects are classified by geographical location, and can be grouped into 4-5 sub-dialects with variation across phonology, morphology, syntax, semantics, and lexicon (Behnstedt and Woidich, 1985; Badawi, 1973). The most prestigious dialect is Cairene Egyptian Arabic (CEA), the sub-dialect spoken by approximately 40% of Egyptians, specifically middle class Egyptians in Cairo and urban cities (Gadalla, 2000; Hanna, 1962; Harrell, 1957; Hospers, 1973; Norlin, 1987; Leddy-Cecere and Schroepfer, 2019). On the other hand, Sa'idi Egyptian Arabic (SEA) is the "the most ridiculed, stigmatised and stereotyped" sub-dialect of EA (Bassiouney, 2017), yet is also the second most spoken EA sub-dialect by 40% of Egyptians. Thus, these sub-dialects are equal in usage but unequal in prestige. EA is the most thoroughly researched DA, yet work on sub-dialects other than CEA is extremely limited. With the exception of Behnstedt and Woidich (1985), and Khalafallah (1969) there exists no recent dialectal surveys on linguistic features of Sa'idi Egyptian Arabic.

2.2 Register Variation

The only standardized and codified writing system of Arabic is MSA (Brustad, 2017; Håland, 2017; Høigilt and Mejdell, 2017). In the past century, EA was "rarely written, and ha[d] little prestige among the people" (Harrell, 1957, p.1). Therefore, there remains no codified written system for EA. It was not until the spread of Social Networking Sites across the past three decades generated a wealth of content written using EA, despite the lack of EA codification (Kindt and Kebede, 2017). Written EA output contains inconsistencies in orthographic representations due to a mixture between using codified MSA as well as developing new orthographic representations for linguistic features exclusive to EA. These features can be dialectal markers; however, the defaulting to MSA in orthographic representation despite different dialectal phonetic representations is exceedingly common. This is a result of influences of standard language ideology and emphasis on 'correctness' in language use (Bassiouney, 2014).

An example is the phonetic representation of the lexical item 'camel'. It is orthographically represented as'جمل'[dʒamal] in MSA, pronounced as [damal] in SEA, and [gamal] in CEA. Speakers of both dialects orthographically represent this word using the codified MSA form, when SEA could represent it to be phonologically reflective of one variation in their sub-dialect'دمل'. Using the codified MSA form is common, making it difficult to detect dialectal markers across Arabic sub-dialects. To our knowledge, there has been no empirical corpus analysis of EA written orthographic patterns across sub-dialects. However, there has been a large effort to identify orthographic patterns in DA written data for the purpose of facilitating and enhancing computational parsing of DA inconsistent orthographic

patterns (Altantawy et al., 2010; Habash et al., 2005; Habash, 2007; Habash et al., 2012; Fashwan and Alansary, 2021). The complexity of Arabic orthography, lack of DA codification, prevalence of MSA as the medium of writing, and lack of empirical research across DA written/spoken registers all motivate the validation of collected DA written text before using this data to represent EA sub-dialects.

2.3 Resources and Tasks

A survey of EA corpora from ***The Linguistic Data Consortium*** (LDC), ***MASADER***[1], and ***InfoGuistics***[2] show that existing Egyptian Arabic corpora primarily feature CEA. MADAR (Bouamor et al., 2018) is a multi-dialect corpus across 25 Arabic cities, one of which is a SEA city. However, MADAR is translated from English and French and not a naturally-occurring corpus, and thus excluded for the purpose of this paper.

DA sub-dialects have been an Arabic NLP focus mostly for dialect identification. A number of other efforts to identify DA sub-dialects on the city level include NADI2020 (Abdul-Mageed et al., 2020a) and NADI2021 (Abdul-Mageed et al., 2021), two series of dialect identification shared tasks. These tasks target micro-dialect identification through matching each Tweet to its corresponding city, with approximately 56 Egyptian cities represented in the datasets. Teams mainly used transformer-based methods for this challenge.

The question is whether the corpora assumed to represent sub-dialects actually do so. The NADI2020 & 2021 sub-dialect shared task's difficulty is reflected in the low F1 scores achieved, 6.39% in NADI2020 and a slight improvement to 8.6% in NADI2021. One reason could be that not all cities have distinct sub-dialects, with some spanning across many cities with minimal distinctions (Behnstedt and Woidich, 1985). Therefore, predicting a specific city is too specific a task when the underlying dialectal features are specific to all cities in the same area. It is also possible that the geo-referenced tweets are not representative of the intended sub-dialects because speakers avoid using less prestigious varieties in certain settings, instead reverting to MSA. This is further explored in this paper.

Abdul-Mageed et al. (2020b) present another contribution towards micro-dialect identification by fine-tuning BiGRU and mBERT models to distinguish sub-dialects in around 21 Arabic countries and 319 cities. They report human annotation at the city level was deemed nearly impossible, as they employed annotators from various Arabic countries to identify sub-dialects outside their native country and dialect. This task would likely be difficult but feasible within a single country. For instance, while a Moroccan might struggle to identify Egyptian sub-dialects across Egyptian cities, an Egyptian might have the linguistic experience necessary to make such distinctions. Despite such annotation efforts, including adjustments for diglossia and code-switching within the data, the system's peak performance was an F1 score of 20.11% and accuracy of 19.88%. The system performed better when utilizing dialectal Arabic alone without inclusion of MSA data. Performance was higher when fewer cities were included.

3 Sub-Dialect Distance Measures

3.1 Dialectal Features

This paper relies on dialectal features from ground-truth dialect surveys to measure the distance between sub-dialects of EA and their expected patterns. Starting with morphological and grammatical features of each sub-dialect, we focus on demonstratives, interrogatives, prepositions, adverbs, and negation particles, as reported for CEA and SEA in existing 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 CEA in orthography, yet are essential to the syntax of SEA and CEA to maximize the likelihood of their presence in the text. Based on the ground-truth surveys, we the believe selected features are sufficient to indicate how well a corpus represents each sub-dialect, although discrepancies in the orthographic representation of these features can vary. For this reason, we rate each feature for markedness; sample features are illustrated in Table 1, and a full list of features in the appendix.

We exclude possible overlap corresponding to MSA features when possible. We use regular expressions to further account for spelling mistakes, such as usage of 'ي، ى' or 'ا ، أ ، إ ، آ' interchangeably, and different orthographic representations 'برضه' vs 'برضو' or 'بردو', and allomorphs of selected features 'ما' vs 'مـ'. We tested features

[1] https://arbml.github.io/masader/

[2] aucegypt.edu/infoguistics/directory/Corpus-Linguistics

in isolation to ensure validity and reliability.

A quantitative analysis of feature validity considers the likelihood of capturing false positives. For example, Ad4 in Figure 2 (left) is able to capture different orthographic representations with less than 5% false positive results. For features with false positives higher than 5%, we analyse them qualitatively. For example, in Figure 2 (right), the SEA negation particle coded Neg4 occurs by adding the suffix 'شي' at the end of a verb. The regex captures this representation, along with any part of speech ending in 'شي', resulting in a large number of false positives, which are then checked manually.

Feature	MSA	CEA	SEA
Interrogative***	أيضاً	برضه	برضك
	'also'	'also'	'also'
Adverb***	الآن	دلوقتي	دلوق
	'now'	'now'	'now'
Particle*	ليس	مش	مش
	'not'	'not'	'not'
Preposition*	في	في	ف
	'in'	'in'	'in'
Demonstrative***	اولائك	دول	داكهما
	'these'	'these'	'these'

Table 1: Sample grammatical features distinctions between MSA, CEA, and SEA. *** indicate most marked features, and * the least marked.

We also elected to exclude features which do not have a clear orthographic distinctions between SEA and CEA from quantitative results. Some features differ between SEA and CEA in phonetic, morphological, or semantic distinctions, however these distinctions are not indicated in the orthographic form. For example, the free negation particle, Neg1, is less likely to be followed by perfective or imperfective verbs in CEA, but this is common within SEA. Accordingly, we elected to examine the data qualitatively with non-orthographically distinct features in order to capture some dialectal features.

3.2 MSA and DA

We measure the usage of MSA in tweets across sub-dialects in one dataset. The more MSA is used in the dataset, the less DA is used and, therefore, the less the sub-dialect is actually represented. To determine usage of MSA, we identify MSA morphological features and isolate tweets, then manually annotate for correctness. Two annotators, native speakers of EA, manually annotate the Micro-Dialect dataset (Abdul-Mageed et al., 2020b) for MSA, DA, and code-switching of both using annotation guidelines for Arabic dialectness by Habash et al. (2008). We group annotation guidelines of 1 & 2 as MSA, 3 as code-switching, and 4 & 5 as DA. Inter-annotator reliability across a sample 1000 tweets measured 86%.

4 Datasets

We use this section to first discuss the baseline or reference corpus which is used to validate the expected features, to determine whether our extraction method does in fact capture the variants which we intend to use to explore sub-dialects of EA. We then describe the corpora used to test whether geo-referenced tweets from specific cities contain the dialectal variants expected given the ground-truth dialect surveys.

4.1 Cairo Baseline Corpus

For the baseline corpus, we use Cairo geo-referenced tweets from Dunn (2020), shown in Table 2. The purpose of this corpus is to ensure the validity of our feature extraction method. Therefore, this corpus is used to measure prevalence of CEA features in Cairo tweets. Tweets include both DA and MSA, and have been pre-processed to only include the Arabic text in the Tweet. With exception of prepositions, CEA features do not overlap with MSA features, therefore, the results should reflect CEA usage in tweets. Due to the size of this corpus, that it is extracted from Cairo, we expect to find high representation of the selected CEA features. Due to migration from SEA cities to Cairo, we also expect to find some SEA features represented by SEA users who might have migrated to Cairo, though much less than its CEA counterparts. Therefore, this corpus is a baseline to ensure the validity and reliability of the script in capturing features by georgraphical location.

Dataset	Tweets	Tokens	MSA/DA
Baseline	808,312	12,233,632	Both

Table 2: Baseline Corpus (Dunn, 2020) across Cairo. Tokens by\ s.

Figure 2: Examples of distinct orthographic representations resulting in false positives (right) vs. alternating orthographic representations of the same word with no false positives (left)

4.2 Sub-Dialect Datasets

We examine three datasets of tweets geo-tagged by city from Arabic micro-dialect identification shared tasks. Datasets include MicroDialect Identification (Abdul-Mageed et al., 2020b), NADI 2020 (Abdul-Mageed et al., 2020a), and NADI 2021 datasets (Abdul-Mageed et al., 2021) across eleven cities. Tweets were collected in 2019 over 10 months, from users who exclusively tweeted from the same location.

Dataset	Tweets	Tokens	MSA/DA
MicroDialect	6,056	77,173	Both
NADI2020	1,021	13,288	Both
NADI2021	798	7,324	DA
Total	7,875	97,785	

Table 3: CEA Datasets: Tweets span across Cairo, New Cairo City, Suez, PortSaid and Ismailia (Abdul-Mageed et al., 2020b; Abdul-Mageed et al., 2020a; Abdul-Mageed et al., 2021). Tokens by\ s.

SEA and CEA cities were determined based on reported dialectal surveys (Behnstedt and Woidich, 1985). Except for NADI2021 (Abdul-Mageed et al., 2021), all tweets include both MSA and DA. All datasets were pre-processed for punctuation, replies, other embedded foreign tokens, hashtags, or indicators for cross-posting on other platforms except for MicroDialect datasets. We elected to not pre-process this corpus to further examine the results on both pre-processed and unprocessed datasets. Some of the original datasets included 10M tweets but cannot be obtained due to API limitations at the time of this paper; therefore, we examine the limited data released within the training and development datasets.

SEADataset	Tweets	Tokens	MSA/DA
MicroDialect	3,076	39,292	Both
NADI2020	1,862	24,693	Both
NADI2021	1,863	16,507	DA
Total	6,801	80,492	

Table 4: SEA Datasets: Tweets span across Qena, Asyut, Aswan, Luxor, Sohag, and BeniSeuf (Abdul-Mageed et al., 2020b; Abdul-Mageed et al., 2020a; Abdul-Mageed et al., 2021). Tokens by\ s.

5 Results

5.1 Does the Baseline Cairo Corpus Contain CEA Features?

To test the validity of sub-dialect morphological CEA and SEA features reported in Behnstedt and Woidich (1985), Khalafallah (1969), and Leddy-Cecere and Schroepfer (2019), we measure the distance between spoken CEA features reported and their prevalence in the written Cairo Baseline corpus. As illustrated in Figure 3, CEA features are overwhelmingly prevalent in the Cairo corpus, while SEA features are not. Each feature is an alternation (i.e., the CEA vs SEA variant). This figure shows the percentage of CEA features used in the baseline Cairo corpus. Feature names correspond to the feature list in the appendix.

This high usage of CEA variants and low usage of SEA variants in the baseline corpus confirms the validity of using these features to measure the distance between dialects. Therefore, we conclude that Cairo is representative of the sub-dialect reported in the dialectal surveys: Cairene Egyptian Arabic. In the next section, we measure the SEA datasets for its representation of SEA sub-dialectal features.

5.2 Do SEA Corpora Contain SEA Features?

The first question is whether we see a greater share of expected SEA features in corpora used to rep-

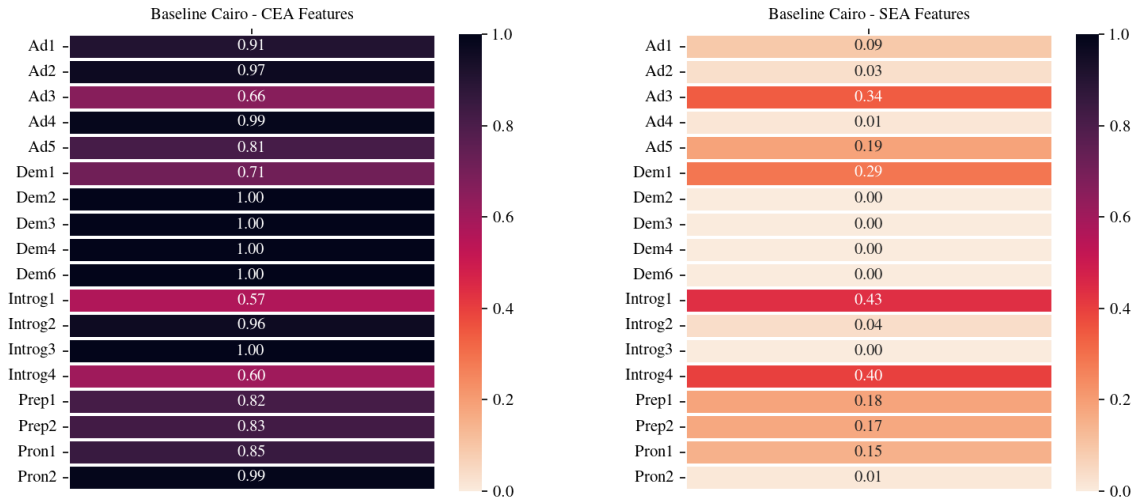

Figure 3: Share of Expected CEA (left) and SEA (right) Variants for each Alternation for Cairo Baseline Corpus. MSA and DA Corpus. Features are complementary.

resent SEA. We take a feature-by-feature look in Figure 4, here using the share of SEA variants for each of the alternations discussed above from the dialect survey. These are complementary features, so that if the share of SEA usage is 25%, then the share of CEA usage must be 75%. Each row is a feature, corresponding with the feature descriptions found in the appendix. The first column represents the baseline corpus of Cairo tweets. The second column represents the CEA cities from NADI2020, NADI 2021, and MicroDialect Corpus and the third column the SEA cities from the same datasets. We would expect, then, that there would be a much higher share of SEA usage in the final column.

First, many features remain unobserved (hence a 0.00 value), even though the annotation methods discussed above accurately identify these variants and some are observed in the Cairo Baseline corpus. This means that the features are simply not observed in these relatively small corpora.

Second, we see that only a few of the overall alternations show the pattern expected from the dialect surveys: Ad3, Dem1, Introg1, Introg4, Prep2, and Pron1 are all markedly more common in SEA as expected. The other features show either no difference at all or the opposite pattern as the dialect surveys. However, what is significant across these specific features is their distinction from their CEA counterparts in either the shortening or elongation of existing vowels or the loss of voiceless final consonants. For example, SEA dialectal surveys report the lack of [h] at the end of Introg4 in SEA features written as ‘لي’ [le:], while CEA surveys report its presence in CEA features written as ‘ليه’ [le:h]. This distinction is not as marked as the distinct realization of Dem6, where SEA features add a stop /k/ at the end of the word written as ‘برضك’ [bardˤak], a phoneme more marked than /h/.

Other highly marked SEA features, such as negation particles, are not observed in SEA datasets yet do occur in the Cairo Baseline corpus. For example, Neg4, Neg3 (Table 5) SEA features are marked with either dropping the CEA negative prefix ‘ما’, and adding a long vowel ‘ي’ to the CEA negation suffix ‘ش’. Qualitative analysis of Neg 3, 4 shows these features are not observed within any SEA dataset, yet are observed in the Cairo Baseline corpus across 50 instances such as ‘ينفعشي’ meaning ‘not possible’, ‘شافشي’ meaning ‘did not see’. This finding extends to other SEA features observed only in the Cairo Baseline corpus, except for Dem 2-6.

Third, overall there is little difference between the CEA and SEA corpora in usage of SEA features. One reason could be language change which has taken place since the dialect surveys. Another reason could be the impacts of internal migration from rural to urban areas (Miller, 2005). If this were the case, then we would expect that some users in CEA cities would maintain clear SEA fea-

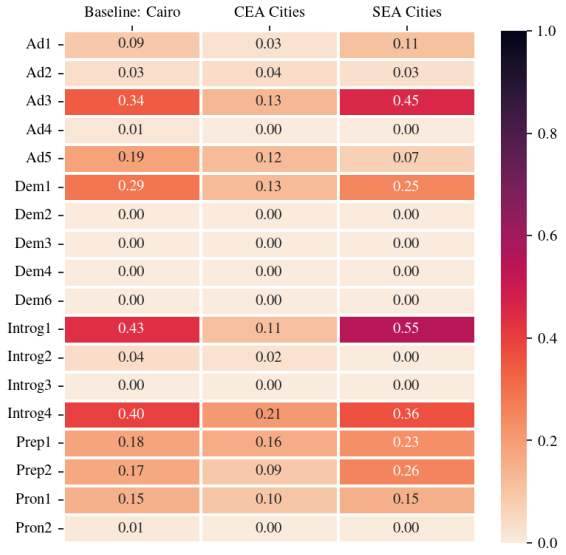

Figure 4: Share of Expected SEA Features for each Alternation across SEA cities and CEA cities in NADI2020, NADI2021, and MicroDialect Corpus compared with baseline Cairo corpus.

tures. This is the goal of the analysis in the next section, where we look at individual cities within each dialect area.

5.3 Are Cities Consistent Within Regions?

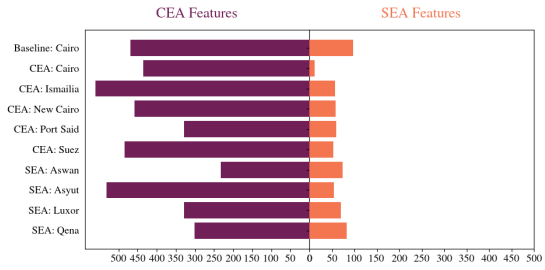

Figure 5: Prevalence of CEA and SEA features by city, using frequency per 10k words. MicroDialect Corpus.

The next question is whether the features expected from the ground-truth dialect surveys appear in the tweets representing different cities within SEA and CEA. While the overall aggregated usage might be unexpected, perhaps some cities have changed (i.e., from SEA to CEA), thus disguising usage in the core SEA cities. This is shown for the mixed MSA and DA Micro-Dialect corpus in Figure 5 & 6, where each city is a bar. The purple values on the left represent the overall frequency

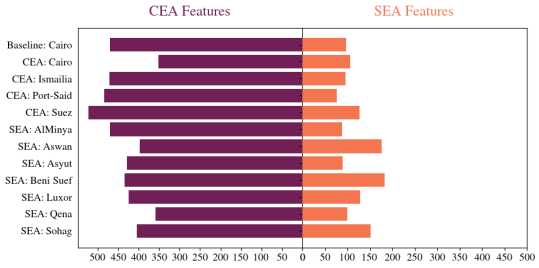

Figure 6: Prevalence of CEA and SEA features by city, using frequency per 10k words. NADI 2020 Corpus.

per 10k words of CEA features (the prestige dialect), and the pink values on the right represent the same quantity for SEA features (the non-prestige dialect).

What we see, first, is that CEA features are overall much more common than SEA features, across both dialect areas. There is a high prevalence of CEA features even in cities expected to represent SEA, such as Asyut, although most SEA cities have a lower rate of usage. Second, we see that there is a relatively equal usage of SEA features across cities, even central CEA locations like Cairo. Because this data represents a small number of users, the figure includes a baseline corpora of other tweets from Cairo, a much larger corpus as described in the Data section. We contrast this raw unprocessed

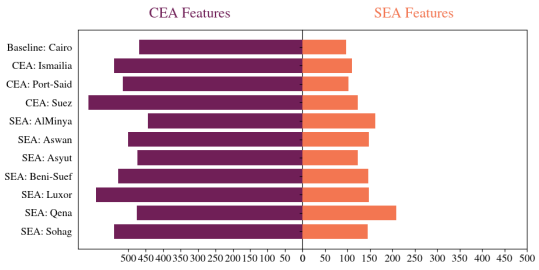

Figure 7: Prevalence of CEA and SEA features by city, using frequency per 10k words. NADI 2021 Corpus.

corpus with the cleaned version in Figure 7, here using the NADI 2021 shared-task corpus. Because MSA samples have been removed here, the overall rate is much higher. However, while the density of dialectal features is higher, there is still no sharp distinction in the usage of CEA features in CEA and SEA locations. On the other hand, SEA features are slightly more frequent in SEA locations. Further, qualitative analysis of Neg1 and Neg2 reflect the results of the quantitative analysis. NADI2021 shows twice Neg1 SEA usage than its CEA counter

part, observing instances of the particle followed by imperfective verbs such as 'مش تستني' meaning 'do not wait' and 'مش فيه' meaning 'there is none'. In CEA, both instances are negated using the Neg2 feature, represented as 'متستنيش' and 'مفيش', respectively. The same analysis for the NADI 2020 corpus is shown in Figure 6. Again the CEA variants dominate across all cities, although to a lesser degree in SEA cities.

5.4 Are Users Consistently Writing in SEA?

The basic finding here is that the corpora representing SEA dialect areas do not contain a substantial usage of the expected SEA features. Instead, CEA features are found across all cities. Why? One possibility is that language change has taken place since the dialect survey was undertaken, although this would be an unusually fast process of change. Another possibility is that older or less connected speakers retain the SEA features but are not represented on social media. A third possibility is that SEA speakers do not produce SEA variants in this digital written setting. We will consider these possibilities further in the discussion.

For now, it is possible that individuals from SEA and CEA cities have changed locations. Thus, we might expect users to consistently use one or the other sub-dialect but to be located in unexpected cities. This is the purpose of the analysis in the next section.

To find out if there are users of each dialect who are out of place, perhaps because of internal migration within Egypt, we visualize the distances between MicroDialect user-specific corpora in Figure 8. Here each point is a corpus representing a single user; the style of each point refers to the dialect area it is supposed to represent. Points are then positioned within a two-dimensional space by using PCA to reduce the usage of all dialectal features into two main components. Taken together, these two components explain 96% of the variance across features; thus, we take this as a reliable visualization of the dialectal relations between user-specific corpora.

First, it is clear that individuals taken to represent both CEA (circles) and SEA (x's) are inter-mingled. This would indicate that the previous overlap in feature usage across CEA and SEA is not because some individuals retain expected usage and others do not. Rather, the usage patterns of individuals are not organized around the expected dialects. In other words, the disconnect between SEA corpora and expected SEA features is not a result of individual differences across users.

Second, since each SEA user is closely patterned with at least one CEA user, this indicates that the core expected SEA speakers are not actually producing that dialect. One possibility is that these users are instead producing either more standard dialectal features (CEA) or are simply reverting to non-dialectal production (MSA). This is explored in the final section.

5.5 Who is Reverting to MSA?

To explore whether SEA users are producing CEA features or resorting to non-dialect production, we annotated the largest SEA corpus, MicroDialect corpus, for MSA, DA, and code-switching. As illustrated in Figure 9, SEA users seem to be using MSA approximately as much their usage of DA. However, CEA users are using DA significantly more than MSA. There were no significant differences in code-switching among both groups.

This tells us two things. First, SEA tweets include a high number of MSA tweets, therefore, chances of SEA feature representation in SEA dataset has lowered by 50% of the overall dataset. Second, when using DA, SEA users do not use highly marked SEA features, but rather resort to either CEA features or SEA features which carry closer resemblance to their CEA counterparts. Regardless, SEA datasets are not representative of the targeted sub-dialect SEA. It is worth noting that the data released is across 11 SEA and 11 CEA users, thereby, limiting any generalizations about SEA data in general. However, insofar as these corpora are taken to represent SEA production, this results show that the non-prestige sub-dialect is inadequately represented compared to the prestige sub-dialect.

6 Discussion

As highlighted by the results, geo-referenced SEA datasets are not representative of SEA sub-dialects. Results are consistent across DA datasets, MSA and DA datasets, and processed and unprocessed datasets. One possibility could be that language change has taken place since the dialect survey was undertaken. However, there is prevalent evidence of SEA features in current SEA speech. Therefore, this cannot be attributed to SEA drastic language change. Another possibility is that older or less

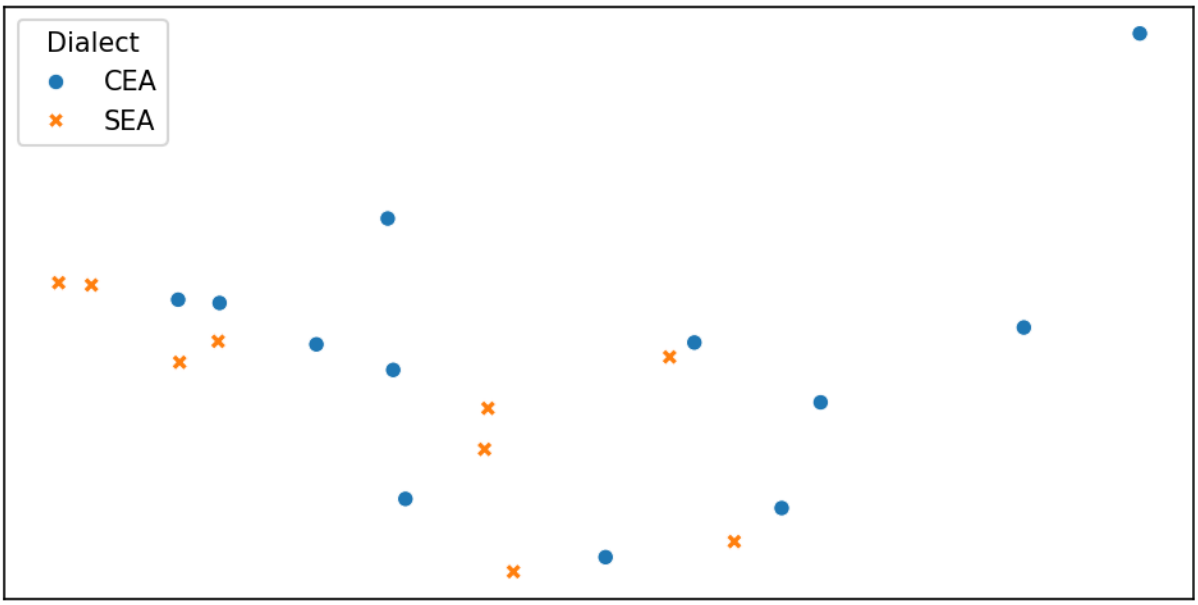

Figure 8: User-by-user plots of feature usage, visualized using PCA for dimension reduction. The original vectors undergoing PCA are the relative frequency of each dialectal feature.

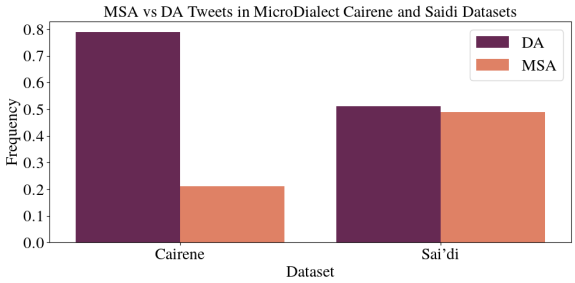

Figure 9: MSA vs. DA Tweet Frequencies in CEA and SEA Datasets.

connected speakers retain the SEA features but are not represented on social media. Kindt and Kebede (2017) report Cairene Egyptians prefer using written MSA vs. DA based on education, age, gender, and platform. Egyptians between the ages of 13-34 use DA significantly more frequently than Egyptians over 50, and women are more likely to write in MSA than men. Given the limited user demographic information beyond consistently tweeting from the same location for over 10 months, we cannot conclude if some, or any, demographic variables contribute to the lack of SEA features or DA use in SEA datasets. A third possibility is that SEA speakers do not produce SEA variants in this digital written setting. While there is a lack of recent SEA dialectal surveys, there is evidence on low attitudes and stigmatized perceptions of the SEA dialect (Bassiouney, 2014; Bassiouney, 2017). SEA users could be avoiding SEA markers in an attempt to position themselves differently across digital platforms. A larger geo-referenced written digital corpus is needed to explore these possibilities further. Regardless, the examined SEA datasets are not representative of the SEA sub-dialect, and register variation is significant across SEA spoken and written registers.

7 Conclusion

This paper finds that EA sub-dialects (except CEA) are low-resourced, and existing Tweet datasets are not representative of EA sub-dialects. Further, register variation between SEA speech and naturally-occurring digital written tweets is significant, therefore, these results call into question the validity of relying on geo-referenced tweets alone to represent dialectal differences. This paper further highlights the need for more representation across DA resources to include DA sub-dialects (Tachicart et al., 2022), and more empirical research on register variation across Dialectal Arabic written sub-dialects and their orthographic patterns in digital spaces.

8 Limitations

Given the inconsistencies across Arabic written DA orthography, the selected morphological markers' orthographic representation is not the ground-truth, but rather the most frequent patterns observed by the authors in EA sub-dialects and DA digital written contexts, in alignment with the dialectal

surveys. This could explain the 0.00 consistent results for some features, although we do account for this through experimenting with all possible orthographic representations. These features could be restricted to speech, or have fallen out of use among the demographic of SEA users online.

Another limitation includes our choice to restrict dialectal markers in quantitative analysis to ones captured with minimal false positives after several iterations of analysis, limiting our quantitative analysis to the most explicit features. We expect some features might be underrepresented or overrepresented in the Cairene Baseline Corpus due its large size, especially if they overlap with similar MSA patterns. We also recognize that some SEA and CEA features are observed in other rural and urban sub-dialects, such as Alexandrian Egyptian Arabic or Shara'wi Egyptian Arabic. However, given the geo-referenced nature of the datasets, we limit our analysis to cities that use CEA and SEA only.

There could be more significant evidence of SEA lexical markers in SEA datasets, however, we do not examine lexical choices between SEA and CEA tweets in this paper. Further, we recognize that one of the most popular Egyptian TV and cinema genres have focused on Sai'di settings (Bassiouney, 2017), and the datasets explored could possibly include quotes or references from such works, and accordingly impact SEA results.

Acknowledgments

We thank Dr. Nizar Habash for the fruitful discussions on representation of Egyptian Arabic subdialects. We are grateful for Dr. Muhammad Abdul-Mageed and his group for compiling and sharing the Micro-Dialect Egyptian dataset used in this paper. Special thanks to Dr. Lane Schwartz for his guidance and valuable input. We also thank the anonymous reviewers for their constructive feedback.

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. *arXiv preprint arXiv:2103.08466*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

Mohamed Altantawy, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological analysis and generation of arabic nouns: A morphemic functional approach. In *Language Resources and Evaluation Conference*.

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.

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

Douglas Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

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.

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

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

Charles A Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Hassan AH Gadalla. 2000. *Comparative morphology of standard and Egyptian Arabic*, volume 5. Lincom Europa Munich.

Nizar Habash. 2007. On arabic transliteration. in van den bosch, a. and soudi, a., editors, arabic computational morphology: Knowledge-based and empirical methods.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Nizar Habash, Owen Rambow, and George Anton Kiraz. 2005. Morphological analysis and generation for arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24.

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.

H Morcos Hanna. 1962. *The phrase structure of Egyptian colloquial Arabic*, volume 35. Walter de Gruyter GmbH & Co KG.

Richard Slade Harrell. 1957. *The phonology of colloquial Egyptian Arabic*. American Council of Learned Societies.

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

Johannes Hendrik Hospers. 1973. *A Basic Bibliography for the Study of the Semitic Languages: Volume I*, volume 1. Brill Archive.

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

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*, 2 edition, pages 433–457. Routledge.

Catherine Miller. 2005. Between accomodation and resistance: Upper egyptian migrants in cairo.

Kjell Norlin. 1987. *A phonetic study of emphasis and vowels in Egyptian Arabic*, volume 30. Lund University.

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

Manfred Woidich. 1996. Rural dialect of egyptian arabic: an overview. *Egypte/Monde Arabe*, (27-28):325–354.

A Appendix - Supplementary Materials

Feature	Gloss	Code	SEA	CEA	MSA
Adverbs	Now	Ad1**	دلوقت	دلوقتي	الآن
Adverbs	Very	Ad2*	واصل، خالص	خالص	كثيراً
Adverbs	Outside	Ad3**	برا	برة، بره	الخارج
Adverbs	Also	Ad4***	برضك، برض	برضه، برضو	أيضاً
Adverbs	Very	Ad5***	قوي	أوي	جداً
Prepositions	On	Prep1*	ع، على	ع، على	على
Prepositions	In	Prep2*	ف، في	في	في
Interrogative	Why	Introg1**	لي	ليه	لماذا
Interrogative	Where	Introg2***	وين	فين	أين
Interrogative	When	Introg3***	ميتى	امتى، امتا	متى
Interrogative	How	Introg4**	كيف	ازاي	كيف
Particles	Negation - free	Neg1*	مش +	مش +	ما، ليس، لا
Particles	Negation- bound	Neg2*	ما+ش	ما+ش	-
Particles	Negation	Neg3**	ما+شي	-	-
Particles	Negation	Neg4***	شي+	-	-
Pronouns	I (am)	Pron1**	آني	أنا	أنا
Pronouns	We (are)	Pron2**	نحنا	احنا	نحن
Demonstratives	This (m.)	Dem1*	دا	ده	هذا
Demonstratives	That (m.)	Dem3**	داك	دا	ذلك، ذاك
	This (f.)	Dem5**	داكهي، دي	دي، ديه	هذه
Demonstratives	That (f.)	Dem6**	داكي	دي، ديه	تلك
Demonstratives	These	Dem7**	داكهما	دول	هؤلاء
Demonstratives	Those	Dem8**	ديكهما	دول	أولائك

Table 5: Sample grammatical features distinctions between SEA, CEA, and MSA. *** indicate most marked features, and * the least marked.

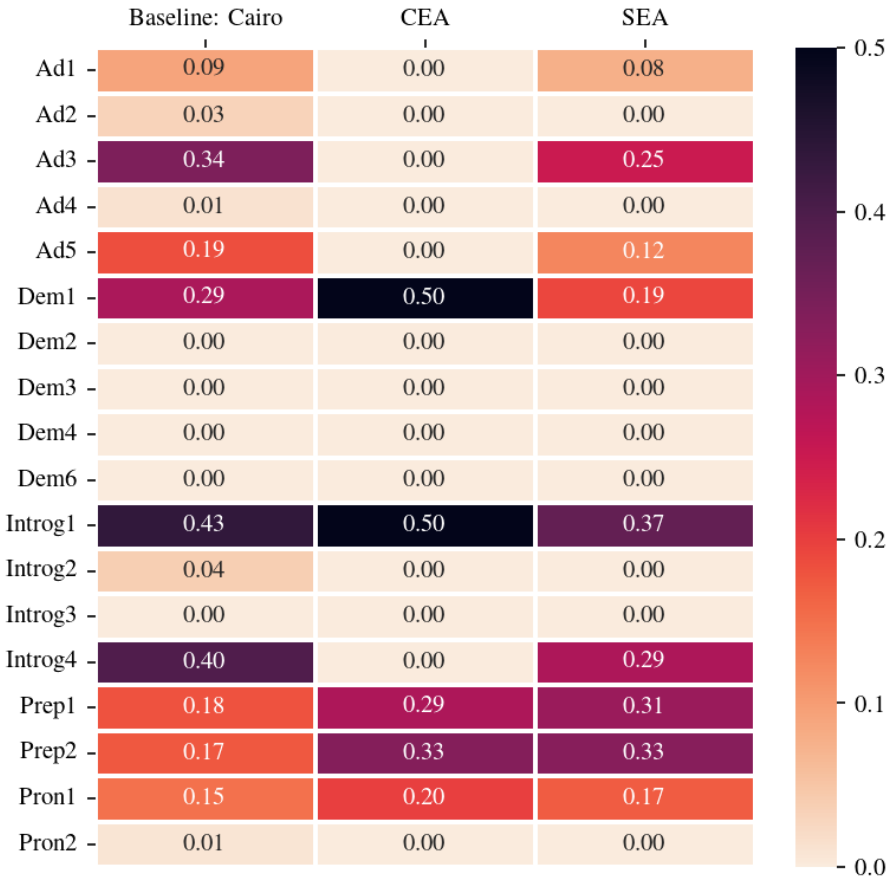

Figure 10: Share of SEA Variants for each Alternation. NADI2020 Corpus compared with Baseline Cairo corpus.

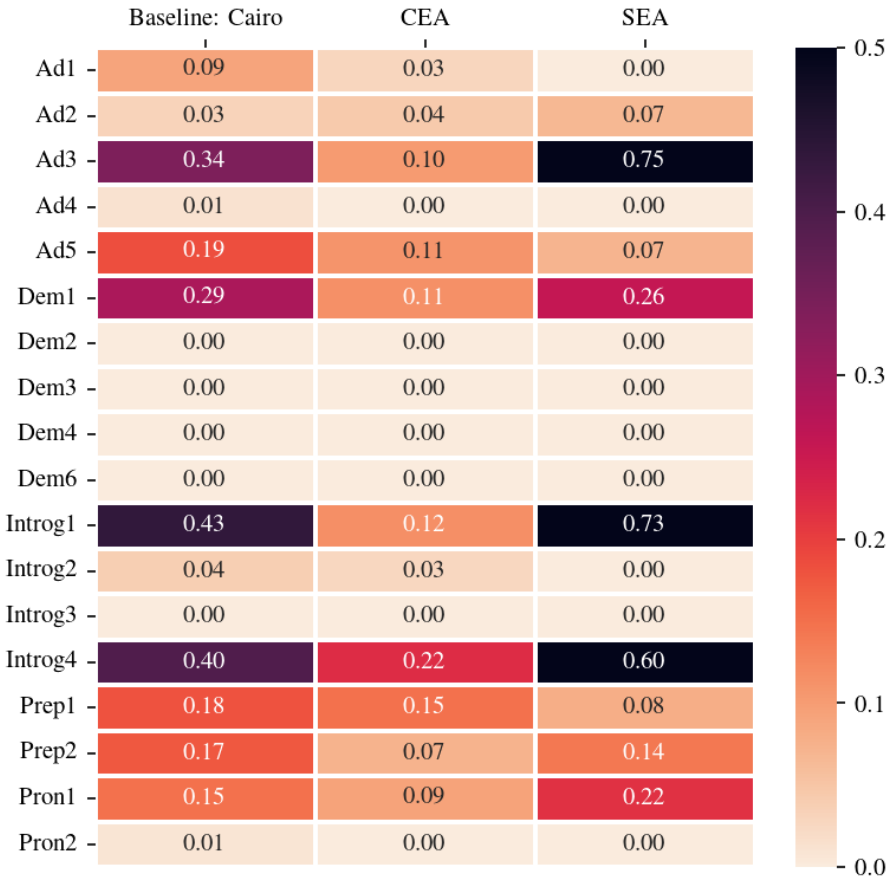

Figure 11: Share of SEA Variants for each Alternation. MicroDialect Corpus compared with Baseline Cairo corpus.

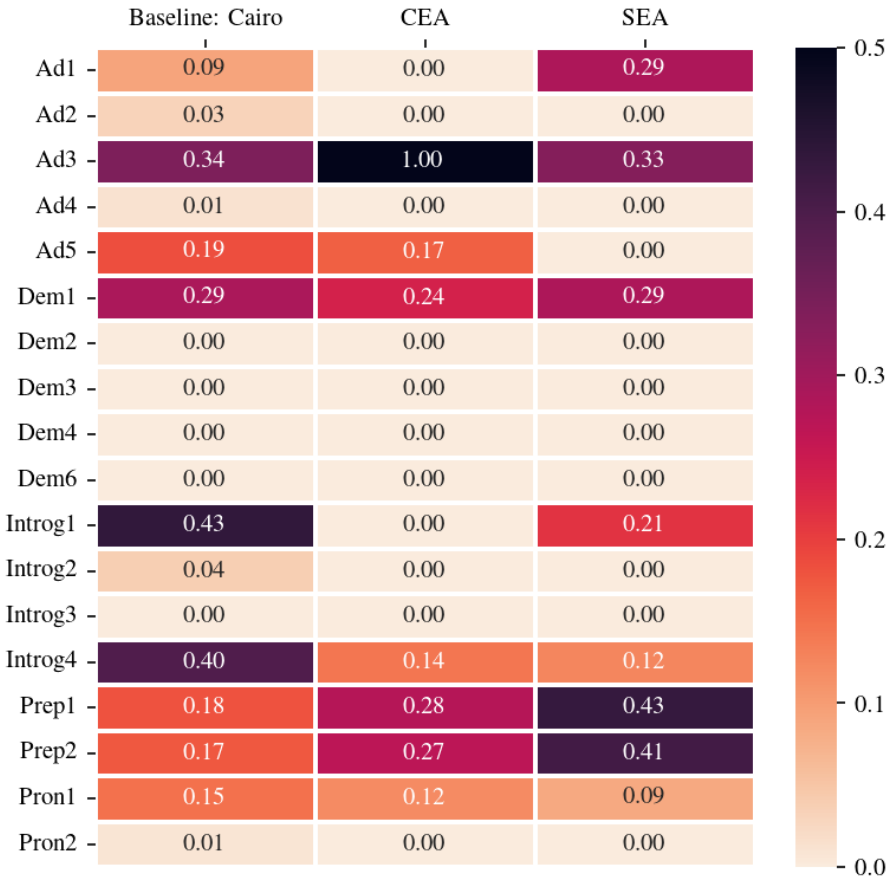

Figure 12: Share of SEA Variants for each Alternation. NADI2021 Corpus compared with Baseline Cairo corpus.