Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4130–4138
Marseille, 11–16 May 2020

A Spelling Correction Corpus for Multiple Arabic Dialects

Fadhl Eryani, Nizar Habash, Houda Bouamor,† Salam Khalifa
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
†Carnegie Mellon University in Qatar
{fadhl.eryani, nizar.habash, salamkhalifa}@nyu.edu, †hbouamor@qatar.cmu.edu

Abstract

Arabic dialects are the non-standard varieties of Arabic commonly spoken – and increasingly written on social media – across the Arab world. Arabic dialects do not have standard orthographies, a challenge for natural language processing applications. In this paper, we present the MADAR CODA Corpus, a collection of 10,000 sentences from five Arabic city dialects (Beirut, Cairo, Doha, Rabat, and Tunis) represented in the Conventional Orthography for Dialectal Arabic (CODA) in parallel with their Raw original form. The sentences come from the Multi-Arabic Dialect Applications and Resources (MADAR) Project and are in parallel across the cities (2,000 sentences from each city). This publicly available resource is intended to support research on spelling correction and text normalization for Arabic dialects. We present results on a bootstrapping technique we use to speed up the CODA annotation, as well as on the degree of similarity across the dialects before and after CODA annotation.

Keywords: Dialects, Corpora, Spelling Correction, Conventional Orthography for Dialectal Arabic

1. Introduction

While the standard form of any language is the variety most likely to receive attention from natural language processing (NLP) researchers and developers, more research is on the rise to address the needs of non-standard varieties and dialects (Zampieri et al., 2019; Bouamor et al., 2019). The Arabic language, spoken by over 400 million people, is in fact a collective of multiple variants, among which Modern Standard Arabic (MSA) is considered the official primarily written variety of education and culture, even though it is not the native language of any speakers. The other variants are known collectively as Dialectal Arabic (DA), but often classified regionally (as Egyptian, North African, Levantine, Gulf, Yemeni) or sub-regionally (i.e, Tunisian, Moroccan, Lebanese, and Qatari). Arabic dialects are the true native languages historically connected to Classical Arabic and many other regional languages. These dialects are primarily spoken, though their dominance on social media is on the rise. Lacking official recognition, they do not have standard orthographies. As a result, dialectal text tends to have a lot of variety and noise (from a computational linguistics point of view). For instance, Habash et al. (2018) reported 27 different spellings for the Egyptian Arabic utterance /mabiʔulhaːʃ/ "he does not say it", that vary in terms of etymological or phonetic spelling decisions.

This high degree of noise is a major challenge for NLP system development as it increases the degree of sparsity in the data. Such noise can be handled using modeling techniques that normalize and cluster variants if DA is the input to the system, e.g. in machine translation from dialects to other languages. However, when the dialect is the target output, as in speech recognition systems (Ali, 2018), or machine translation into the dialects (Erdmann et al., 2017), evaluation and thus optimization may struggle.

A number of efforts in Arabic NLP have argued for the creation of a common convention for Arabic dialect spelling, named Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2012; Jarrar et al., 2014; Zribi et al., 2014; Saadane and Habash, 2015; Khalifa et al., 2016; Habash et al., 2018). The majority of resources involving CODA annotation consider it a side task to efforts like morphological disambiguation, diacritization and lemmatization, as opposed to being the main target task (CODA *for* CODA).

In this paper, we explore and report on the task of CODA annotation, i.e., spelling correction into the CODA convention.[1] We work with a unique corpus of parallel multiple Arabic dialects, the MADAR Corpus (Bouamor et al., 2018), focusing on five cities: Beirut, Cairo, Doha, Rabat and Tunis.

Our contributions are threefold. First, we created a parallel CODA version of a parallel multi-dialectal corpus, a unique resource, first of its kind. Second, we describe and follow a bootstrapping technique for CODA creation, and we report on its speed and initial accuracy under different pre-existing resource settings. Finally, we quantify the degrees of similarity across the dialects we work on using the annotated data in both Raw and CODA spaces. As expected CODA reduces the overall vocabulary within dialects and increases the overlap across them. The corpus will be publicly available for research purposes.[2]

In the rest of this paper, Section 2 presents some related work. Section 3 introduces the CODA conventions. Sections 4 and 5 discuss our approach and results, respectively. We conclude and present some future directions in Section 6.

[1] While we recognize that the term "spelling correction" evokes a claim of an "official standard," we observe that there are no authorities interested in creating such a standard in the Arab world. And given the growing number of NLP papers and tools working with CODA, it is slowly becoming the de facto standard, at least for NLP. Finally, for the sake of clarity of purpose, we find the term "spelling correction" in a NLP context clearer than "spelling conventionalization".

[2] `http://resources.camel-lab.com/`

2. Related Work

Automatic DA processing has been attracting a considerable amount of research in NLP (Shoufan and Al-Ameri, 2015), facilitated by the newly developed monolingual and multilingual dialectal corpora. Several mono-dialectal corpora covering different Arabic dialects at different granularity levels (region, country and city levels) were built and made available (McNeil and Faiza, 2011; Zaidan and Callison-Burch, 2011; Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Salama et al., 2014; Jeblee et al., 2014; Al-Badrashiny and Diab, 2016; Zaghouani and Charfi, 2018; Abdul-Mageed et al., 2018).

As for dialect-to-dialect parallel corpora, Bouamor et al. (2018) presented the MADAR Corpus, a large-scale collection of parallel sentences covering the dialects of 25 Arab cities alongside the English, French and MSA parallel texts. This resource was a commissioned translation of the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) sentences from English and French to the different dialects. It includes two corpora. The first corpus (Corpus-26) consists of 2,000 sentences translated into 25 Arab city dialects in parallel. The second corpus (Corpus-6) has 10,000 additional sentences (non-overlapping with the 2,000 sentences) from the BTEC corpus translated to the dialects of only five selected cities: Beirut, Cairo, Doha, Rabat and Tunis. The translators, identified from each of the 25 cities specifically, were asked to read a set of sentences provided in English or French and produce a natural translation in Arabic script that precisely reflects the source sentence without any guidance on the orthography.

In all of the above-mentioned corpora, texts are written without following any spelling conventions or standards, which are necessary for building efficient NLP tools and applications. To alleviate this bottleneck, several efforts have been introduced to modernize and extend Arabic orthography and develop orthographic conventions for Arabic dialects. Habash et al. (2012) introduced the concept of Conventional Orthography for Dialectal Arabic (CODA), the very first effort to present a set of guidelines and exception lists for Egyptian Arabic orthography.

Although the first CODA was developed for Egyptian Arabic, it was designed with extensibility in mind. As Egyptian CODA began to be integrated into several Egyptian Arabic resources (Maamouri et al., 2014; Diab et al., 2014; Pasha et al., 2014; Eskander et al., 2013; Al-Badrashiny et al., 2014), other efforts began to extend CODA's coverage into new dialects. 2014 saw the creation of two additional guidelines, Tunisian CODA (Zribi et al., 2014) and Palestinian CODA (Jarrar et al., 2014). Using a variant of CODA adopted for speech recognition, Ali et al. (2014) demonstrated reduced out of vocabulary (OOV) and perplexity for texts rendered in CODA. More dialects have followed since then, with the creation of Algerian CODA (Saadane and Habash, 2015), Moroccan CODA and Yemeni CODA(Al-Shargi et al., 2016), and Gulf CODA (Khalifa et al., 2018). More recently, CODA has garnered the interest of literacy, pedagogy, and heritage specialists as a convenient orthographic standard, such as a website that teaches Palestinian Arabic,[3] amongst others. These efforts were unified in overall principles, namely in how to spell open class words. But during creation of these CODA extensions, each dialect tended to curate its own list of exceptional spellings for closed class words. With the growing number of dialects being incorporated, Habash et al. (2018) presented a more Unified Guidelines and Resources for Arabic Dialect Orthography — dubbed CODA* (CODA-Star) as in for any dialect — specifying closed class spelling in more detail and unifying the CODA creation process. CODA* has since been used to represent over two dozen Arabic dialects.

It is worth noting that in recent years, the problem of spell checking and spelling error correction for Arabic has been investigated in a number of research effort (Attia et al., 2016; Watson et al., 2018). The QALB (Qatar Arabic Language Bank) project (Zaghouani et al., 2014) aimed at building an annotated corpus of manually corrected MSA text for building automatic correction tools, and it was used in two shared tasks on MSA spelling correction (Mohit et al., 2014; Rozovskaya et al., 2015).

3. CODA: Conventional Orthography for Dialectal Arabic

3.1. The Orthography of Arabic and its Dialects

As mentioned in the introduction, Arabic is a family of variants, among which MSA is the official standard language. However, MSA is not the native language of any speakers of Arabic. In unscripted situations where spoken MSA would typically be required (such as talk shows on TV), speakers usually resort to repeated code-switching between their dialects and MSA (Abu-Melhim, 1991; Bassiouney, 2009). Arabic dialects vary phonologically, lexically, and morphologically from MSA and from each other; and they vary from region to region and to a lesser extent, from city to city in each region (Watson, 2007). While MSA has a well-defined standard orthography, Arabic dialects have no official orthographies. Usually, people write in a way that reflects the phonology or the etymology of the words. As such, besides unintentional typographic errors, no spelling of a dialectal word can be considered truly "incorrect." Following (Eskander et al., 2013), we refer to this as *spontaneous orthography*.

Table 1 presents several examples of the degree of variety in dialectal spelling in each of the five dialects in our corpus. For instance, the word استاذ *AstAð*[4] 'professor' in Beirut was written in four different ways reflecting the phonological difference in pronouncing the word in Levantine Arabic (/ʔisteːz/) from MSA (/ʔustaːð/) in some cases, and maintaining the etymological relation with the MSA by spelling the word as if it is pronounced in MSA in other cases.

3.2. CODA Overview

CODA* (pronounced 'CODA star', as in for any dialect) is a conventional orthography for dialectal Arabic presented

[3] http://www.learnpalestinianarabic.com

[4] Arabic script transliteration is presented in the one-to-one Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007). Phonological forms are presented in IPA.

Dialect	Pronunciation	English	Variations	
Beirut	/halla?/	'now'	{ هلق ، هلأ ، هلا }	{ *hlq , hlÂ , hlA* }
	/?iste:z/	'professor'	{استاذ ، استاز ، أستاذ ، إستاز}	{*AstAð , AstAz , ÂstAð , ĂstAz*}
Cairo	/barra/	'outside'	{ برا ، بره ، برة }	{ *brA , brh , brħ* }
	/?innaharda/	'today'	{النهارده ، إنهاردة ، انهاردة ، انهاردا}	{ *AlnhArdh , ĂnhArdħ , AnhArdħ , AnhArdA* }
Doha	/haði:tʃ/	'this'	{ هذيج ، هذيك }	{ *hðyj , hðyk* }
	/wa:jid/	'very'	{ وايد ، واجد }	{*wAyd , wAjd* }
Rabat	/bla:sˁa/, /bla:sˁet/	'place', 'place of'	{ بلاصا ، بلاصت ، بلاصه ، بلاصة }	{ *blASA , blASt , blASh , blASħ* }
	/nta/	'you'	{ نتا ، انتا ، أنت }	{ *ntA , AntA , ÂntA* }
Tunis	/ʃniyya/	'what'	{ شنيا ، شنية ، شنيه ، اشنيه }	{*šnyA , šnyħ , šnyh , Ašnyh*}
	/barʃa/	'very'	{برشا ، برشه ، برشى }	{ *bršA , bršh , bršý* }

Table 1: Examples of spelling variations of the same word in each dialect, as they appear in our CODA annotated corpus.

by Habash et al. (2018). CODA* builds on and unifies a number of previous dialect specific CODA conventions (Habash et al., 2012; Zribi et al., 2014; Jarrar et al., 2014; Khalifa et al., 2018). Since we do not deal with any of the previous dialect specific efforts here, we refer to CODA* simply as CODA in the remaining of this paper. CODA is designed primarily for the purpose of developing computational models of Arabic dialects. For generating our corpus, we follow Habash et al. (2018)'s latest guidelines and resources.[5] Next we go over a few high-level observations about CODA and the CODA creation process pertinent to the results discussion in Section 5.

As mentioned above, spontaneous orthography tends to reflect the etymological or phonological reference a writer may ascribe to a word. In this sense, CODA strives to regulate some of these natural spelling tendencies in an internally consistent system and (generally) according to a MSA reference, more or less familiar to everyone. As Habash et al. (2018) explain, CODA's design tries to "strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities," following a sense that the success of such optimization would ensure CODA stays easily learnable and seamlessly readable to the average Arabic speaker without compromising their ability to interpret a written form in their own dialect.

Sounds and letters The phonetic inventory of DA can vary significantly from one dialect to another. One of CODA's principal insights is in organizing the most common of these changes and linking them to their MSA root cognates. This allows for a word like قمره *qmrh* 'his moon', MSA (/qamaruhu/), to be simultaneously correctly interpreted as (/?amaru/) and (/gamara/) in Egyptian Arabic and Gulf Arabic, respectively. This is because any Arab speaker with minimal knowledge of MSA will recognize how a root radical, in this case ق *q*, is pronounced in their dialect as opposed to MSA. On the other hand, phonological spelling, e.g., امره *Amrh* might be interpreted correctly by Egyptian speakers, but not Gulf speakers.

Morphology Another area in which spontaneous orthography varies is in the choice to cliticize or split certain morphemes, e.g., particles and indirect objects. Following MSA, CODA always splits indirect objects and always spells single-letter clitics attached while separating multi-letter ones (except for the ال *Al* determiner).

The determiner is always spelled out morphemically, notwithstanding coronal assimilation with so-called Sun Letters (Habash, 2010). In its preference for morphemic spelling, CODA also differentiates between similar sounding morphemes such as the pronominal 3rd person singular clitic ه *h* and the non-first person plural verbal suffix وا *wA*, both of which are often rendered in spontaneous text as a word final و *w*. Using our previous example but focusing on Egyptian Arabic, the utterance (/?amaru/) may be rendered امرو *Amrw* in a phonologically inspired spontaneous orthography. Outside of a larger context, this creates triple ambiguity in the token as قمره *qmrh* 'his moon', امره *Amrh* 'he ordered him', or امروا *AmrwA* 'they ordered'. Context plays a key role in resolving such cases.

Closed Class Words CODA guidelines detail specific rules for different categories of closed class words such as numbers or demonstrative pronouns. Numbers for instance create a disproportionate amount of variants and are therefore normalized more aggressively than other classes of words, whereas demonstratives generally keep their phonological spelling.

Finally, a quick note on some trivial CODA rules that can be applied automatically. Punctuation for instance always concatenates to the preceding token. The other case involves word initial Alif-Hamza forms أ, إ *Â, Ă* which in CODA are always spelled as bare Alif ا *A*.

Spelling in CODA As shown in this overview of CODA, generating a CODA spelling can be an involved process requiring careful observation of linguistic facts of the phonology, morphology, and meaning of an utterance, as well as knowledge of CODA's list of frequent root cognate mappings, and other rules. Because of the opaque nature of unvocalized Arabic orthography, some linguistic facts become hard if not impossible to discern at the word token level, even in standard form. Moreover, vowel quality and length vary significantly among dialects, and discerning them requires an annotator that is familiar with how they tend to be realized in that particular dialect. The CODA guidelines include the CODA Seedlex, a convenient reference for looking up CODA spelling decisions.

[5] http://coda.camel-lab.com/

sentence #	word #			
411	EN	They still look a bit green.		
411	FR	Ils m'ont l'air encore un peu débutants.		
==========		**Raw**	**CODA'**	**CODA**
411	1	يظهرلي	يظهرلي	يظهر لي
411	2	مازالو	ما زالوا	ما زالوا
411	3	مبتدئين	مبتدئين	مبتدئين
411	4	شويا	شوية	شوية
411	5	.#	.#	.#

Figure 1: An example of CODA annotation of a sentence extracted from the Tunisian side of the corpus, along with its English and French equivalents. Word 1 shows an example of word splitting in CODA. Word 2 shows an example of both splitting and substitution through final letter addition. Word 4 shows an example of final letter substitution.

4. Approach

4.1. Corpus Selection

In this paper, we focus on five city dialects: Beirut, Cairo, Doha, Rabat and Tunis. We work with 2,000 sentences for each dialect from the MADAR Corpus-26, described in Section 2.

4.2. Annotation Guidelines and Quality Control

While most CODA-annotated data has been created as part of larger morphological and syntactic annotation efforts, the work we present here is unique in that it is strictly focused on CODA. Following Habash et al. (2018)'s latest CODA* guidelines, a native Arabic speaker familiar with a number of dialects was tasked with carrying out the annotations. The annotator worked closely with native informants, particularly for some of the less familiar usages. Beyond CODA annotation, this task was done with an eye towards streamlining the annotation process, extending CODA's dialectal coverage, and facilitating future plans to carry this task out on a larger scale.

4.3. Annotation Process

Manual annotation was done using a Google Sheet setup like the one illustrated in Figure 1 presenting an annotated Tunisian sentence. We use the term Raw sentence to refer to the original text as is. With a Raw sentence as input, a simple tokenizer splits the sentence by white-space and separates punctuation. We refer to each token in the tokenized sentence as the Raw token. Separated punctuation is attached to a "#" symbol marking the direction for concatenation once the sentence is put back together. Similarly, annotators use "#" to annotate concatenation edits.

At the basic level, the task of the annotator is to read each Raw token and type its corresponding CODA compliant spelling. We describe types of edits and their frequencies in the next section. These annotations are done within the context of the sentence. For reference, parallel translations of the sentence are provided to help clarify and disambiguate. As an ongoing effort, CODA guidelines and resources are periodically updated as more dialectal data is handled. Another aspect to the task of the annotator is to mark new linguistic phenomena to be integrated into the CODA guidelines, particularly for words whose CODA spelling is not sufficiently clarified. As a result of this effort, CODA guidelines and resources are being extended with many additional closed class words from Tunis and Rabat, for instance.

4.4. Annotation Speed Up

In order to enhance speed and accuracy, we bootstrap the annotation process by priming annotators with semi-automatically generated CODA suggestions. CODA' (CODA *prime*) denotes any likely Raw to CODA mapping stripped of their original context. The several steps to this process boil down to two tricks. The first step is to leverage previous annotations to create a likelihood estimate model $P(CODA|Raw)$ that ranks likely CODA tokens given a particular Raw token by frequency. In ambiguous cases, such as when multiple viable CODA spellings map from a particular Raw token, less frequent mappings can be displayed as secondary options, allowing the annotator to pick the the best fitting annotation for a particular context. The second step takes care of the out of vocabulary (OOV) tokens unseen in our previous annotations, whose unique set is gathered separately and annotated out of context before they are used to supplement CODA' suggestions. In this step, annotators are asked to produce a single CODA' suggestion based on what they think is most likely. Given that every token has to be manually verified again **in context**, we filter out singletons from the OOV annotation to avoid redundant checking.

At the basic level, annotation speed is limited by the annotators reading speed and the number of edits required. In this sense, annotation speed should be correlated with the accuracy of the CODA' prediction, allowing the annotator to simply copy and paste the right answer with minimal manual edits. Since the frequency of each annotation is recorded, this adds the benefit of making sure CODA rules are applied uniformly, allowing the annotator to flag alternative CODA' suggestions either as viable alternatives in different use contexts, or as annotation errors that need to be re-annotated. We report on speed and accuracy using varying sizes of training data in Section 5.

Figure 2 shows a basic illustration of the bootstrap process. The first step is to load any prior existing annotations into a frequency dictionary of Raw to CODA pairs. We can then extract the OOV terms present in the new text that were

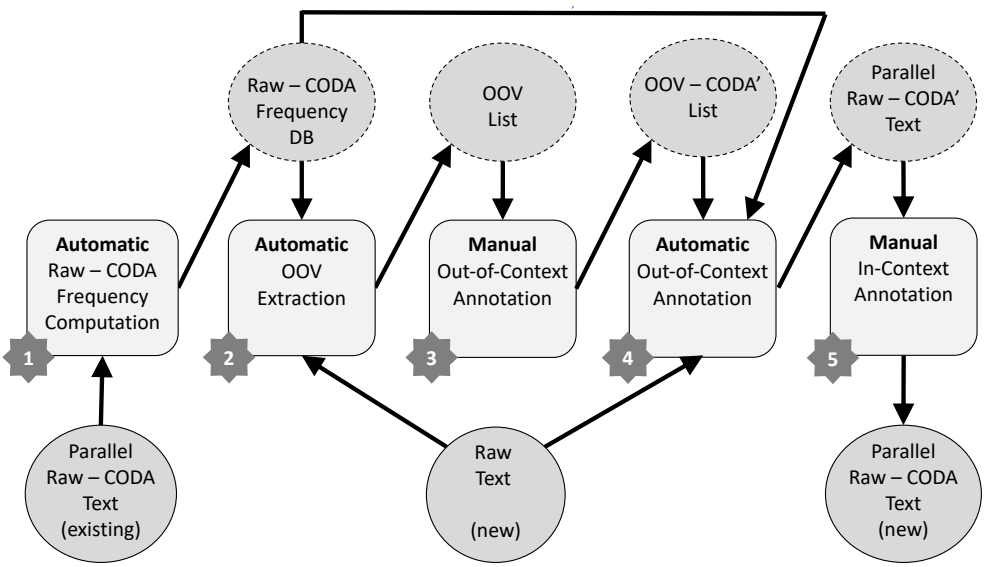

Figure 2: The CODA annotation bootstrapping process

not encountered in prior annotations in step 2 and annotate them for CODA' in step 3. Step 4 involves linking the Raw text with the dictionary suggestions and annotated OOV list to generate CODA' suggestions, which can finally be validated in context. Once this annotation is complete it can be used to extend training data for future CODA annotations. CODA bootstrapping code will be made available along with the corpus. A future implementation of this process can involve periodic or real time consistency validation, using previous annotations to inform future ones on the go, and an option to turn off unwanted suggestions that clutter the suggestion space. The process can also integrate more sophisticated language models to generate predictions.

5. Annotation Results

In this section, we present some analysis of our corpus. We examine the relationship between the Raw and the CODA parallel texts for each city dialect separately. Then, we compare vocabulary coverage across the different dialects and MSA. We also present some results on a speed analysis of the annotation method we used. Finally, we present a learning curve study on an MLE model for automatic CODA annotation trained on data from the same dialect as well as from a multi-dialectal mixture.

5.1. Mono-Dialectal Corpus Analysis

	Tokens		Token/Sentence	
	Raw	**Coda**	**Raw**	**Coda**
Beirut	13,406	13,370	6.7	6.7
Cairo	14,484	14,464	7.2	7.2
Doha	13,359	13,354	6.7	6.7
Tunis	13,879	13,894	6.9	6.9
Rabat	14,944	14,802	7.5	7.4
Average	**14,014.4**	**13,976.8**	**7.0**	**7.0**

Table 2: Corpus token and sentence statistics

	No Edit	Sub	Split	Del
Beirut	81.28%	17.35%	1.38%	0.00%
Cairo	85.98%	12.44%	1.54%	0.03%
Doha	94.90%	4.79%	0.30%	0.01%
Tunis	85.48%	12.66%	1.79%	0.07%
Rabat	83.66%	14.67%	1.66%	0.01%
Average	**86.26%**	**12.38%**	**1.33%**	**0.02%**

Table 3: Raw-to-CODA edit statistics in token space

Corpus Edit Statistics For the average dialect, 86% of tokens tend to be in a CODA compliant form already.[6] With an error rate of just 5%, the Doha set stands out as the most effortless to spell correctly in CODA. Other dialects contained between 14% (Cairo) to 19% (Beirut) non-CODA compliant spellings.

In Table 3, we report three types of edits between each CODA and its parallel Raw form. Splits are defined as white space insertions that separate a token into several. Substitutions (Sub) are any change to the token that is not a split. Included in the substitution count are merges, coded with an insertion of a "#" symbols at either end of a token to mark concatenation with the adjacent token, such as when a single character particle is spelled separate from its base word. Repetition typos constituted the small number of deletions (Del), coded by annotators as "del".

Of the portion of Raw text containing errors – about 14% of the average dialect – substitutions and splits average around 90% and 10% respectively. The most frequent splits involve the separation of indirect objects and multi-character particles. Substitutions involve a wide variety of decisions, such as the root cognate replacements described in Section 3.

[6]Trivial corrections, specifically punctuation and word-initial Alif normalizations are mainly handled automatically and are thus excluded from these calculations.

	Raw Types	CODA Types	Type Overlap	Vocabulary Reduction
Beirut	4,114	3,877	80%	6%
Cairo	4,114	3,820	84%	7%
Doha	3,417	3,377	94%	1%
Tunis	4,044	3,834	84%	5%
Rabat	4,225	4,000	83%	5%
Average	**3,983**	**3,782**	**85%**	**5%**

Table 4: Vocabulary size in number of types and type overlap between the Raw and CODA corpora for each dialect

In the following sections, we compare the the Raw text and its parallel CODA to study the effect of these transformations on the number of tokens in the corpus as well as the size of its vocabulary.

Corpus Token Statistics Table 2 shows the number of tokens for each dialect in Raw and in CODA space. On average, there is almost no change in the total number of tokens (except for a small 0.3% reduction in the CODA set), nor is there a notable change in the number of tokens per sentence.

The difference between spontaneous orthography and CODA in terms of token count is negligible given the small number of splits and concatenations that differentiate Raw from CODA. Because CODA concatenates all single letter particles, a Beirut utterance such as (/ʕ+al+bayt/) 'to the house' is often rendered ع البيت ʕ *Albyt* in spontaneous orthography, but concatenates in CODA to عالبيت ʕ*Albyt*. The choice to not attach the preposition ع ʕ 'on/over/toward' reflects the spelling of its longer cognate in MSA: على ʕ*alaý* 'on/over'. On the other hand, CODA also splits all indirect objects into separate tokens, balancing out the count for concatenations.

Corpus Type Statistics Table 4 presents the reduction in number of unique types when Raw text is rendered in CODA. An average reduction of 5% shows a significant decrease in number of types that is not reflected in token space, suggesting that about 5% of the Raw data's vocabulary contained noisy variants that CODA normalized.

Reaffirming this suggestion is the fact that many Raw types map to the same CODA, at a rate of 1.06 to 1.0. On the flip side, the number of CODA types that map from a single Raw type is 1.01, an indication of instances where ambiguity is resolved, as in the example in the morphology discussion of Section 3.

The reduction in word overlap between the Raw and the CODA texts appears to be correlated with the reduction in vocabulary, as well as with the the amount of edits required for each of the dialectal sets as shown in Table 3. If all what these substitutions did was unify variants, there would be no decrease in word overlap between the Raw and CODA. The fact that 15% of the CODA vocabulary does not overlap with its parallel Raw text shows that much of the CODA vocabulary in our corpus did not resemble any variants that were encountered in the Raw text. This is not surprising given the small size of our corpus.

5.2. Cross-Dialectal Corpus Analysis

In this section, we compare the vocabulary of the parallel texts from different cities using a vocabulary overlap measure.[7] The results of this analysis are presented in Table 5. The upper half of Table 5 (labeled (a) Raw) presents the vocabulary overlap over the original Raw text between each possible city pair; while the lower half of Table 5 (labeled (b) CODA) does the same over the CODA version of the text. The table also computes vocabulary overlap of the various cities against MSA, against **all** other city dialects (e.g., Beirut vs Rabat+Tunis+Doha+Cairo) with and without MSA.

Counting over Raw text types, the average overlap between pairs of cities is about 30%, with minimum overlap of 25% between (Beirut-Rabat) to a maximum of 39% for (Doha-Cairo). The average overlap between these city dialects and MSA is less, at about 28% with a minimum of 23% for (Tunis-MSA) and a max of 36% for (Doha-MSA).

In terms of both the Raw and CODA corpora, the average overlap between any of the sets is close to one third, though it is slightly higher between dialects than between MSA and any of the dialects. The CODA corpus for its part magnifies cross-dialectal overlaps by an average of 6%, while raising the overlap of any one dialect with MSA by only 1%.

Finally, the average overlap of any dialect with the union of the other dialects (with and without MSA) is about 52% and 55%, respectively. This indicates that similarity across dialects is complementary. Roughly speaking, a third of a dialect's vocabulary in our data is unique to itself, while another third can be seen as similar to a specific other dialect, and the rest is distributed amongst all other dialects.

5.3. Manual Annotation Speed Analysis

We measured annotation speed under controlled experimental settings and report them in Table 4. The experiment was divided into sets of a 100 sentences. Each set was annotated under varying bootstrap settings. The first trial is a line-by-line annotation without any bootstrapping.

The second trial shows the effect of the OOV bootrstrapping method (step 3 in Figure 1). Two minutes of out-of-context annotation cut the annotation time by 12 minutes, an increase in annotation speed corresponding to 20% and 30% in the word per minute and sentence per minute rates respectively. This step is carried out in each subsequent trial.

The third trial trains suggestions on a 100 previously annotated sentences, cutting time by another 30%. Quadrupling the size of the training data shows a steady increase in annotation speed. With only 1,600 training sentences, we more than doubled the speed (word/min).

5.4. Automatic Annotation Accuracy

Table 7 shows the accuracy of the MLE baseline with varying sizes of training data.[8] With a total of 2000 sentences

[7] Vocabulary overlap is the number of unique words in the intersection of two sets of unique words from two dialects divided by the smaller of the two sets.

[8] We exclude punctuation normalization from this particular calculation since they are trivial and tend to over-inflate the accuracy.

(a) Raw							
	Cairo	**Doha**	**Tunis**	**Rabat**	**MSA**	**ALL DIA**	**ALL DIA+MSA**
Beirut	34%	37%	26%	25%	26%	51%	54%
Cairo		39%	27%	26%	31%	54%	59%
Doha			30%	29%	36%	59%	64%
Tunis				31%	23%	48%	51%
Rabat					24%	46%	49%
Average	**30%**				29%	53%	57%
(b) CODA							
	Cairo	**Doha**	**Tunis**	**Rabat**	**MSA**	**ALL DIA**	**ALL DIA+MSA**
Beirut	41%	44%	31%	29%	28%	59%	61%
Cairo		45%	32%	31%	33%	62%	67%
Doha			34%	32%	35%	65%	69%
Tunis				36%	24%	55%	57%
Rabat					24%	51%	54%
Average	**36%**				30%	60%	63%

Table 5: Cross-dialectal vocabulary overlap

	Manual Annotation Time (min)				
Training Size (Sen)	**Out-of-context**		**In-context**	**Total**	**word/min**
0	no	0.0	42.4	42.4	17.8
0	yes	2.3	30.4	32.7	20.9
100	yes	0.9	28.0	29.0	27.0
400	yes	0.7	22.2	22.9	31.9
1600	yes	0.1	16.8	16.9	39.7

Table 6: Increase in annotation speed using bootstrapping techniques. The first row shows speed without any bootstrapping. The second row shows the effect of using annotated OOV suggestions without training on pre-annotated data. Rows 3-5 show the effect of using varying numbers of pre-annotated sentences to train suggestions along with OOV bootstrapping.

Training Data															
Other Dialects	**0**					**4000**					**8000**				
Within Dialect	**0**	**125**	**250**	**500**	**1000**	**0**	**125**	**250**	**500**	**1000**	**0**	**125**	**250**	**500**	**1000**
Beirut	80	87	88	91	92	86	89	90	91	92	86	89	91	91	92
Cairo	83	89	91	92	92	88	91	92	93	93	88	91	92	93	93
Doha	93	95	95	96	96	95	95	95	96	96	94	95	95	96	96
Tunis	84	88	89	91	92	89	91	91	93	93	89	91	92	93	93
Rabat	79	84	85	88	89	85	87	88	89	90	86	87	88	89	90
Average	85	89	90	92	92	89	91	92	92	93	89	91	92	93	93

Table 7: Accuracy Results in (%), with varying number of sentences for mono and multi-dialectal training data. The two top row headers (Other Dialects and Within Dialect) specify the training data composition in terms of numbers of sentences and dialect mix.

per dialect, the first 1,000 sentences within each dialect were used to train the MLE model and tested against the remaining 1,000 sentences. On average, an untrained "do nothing" baseline achieves an 85% accuracy, with a maximum of 93% for Doha and a minimum of 79% for Rabat. Training on just 125 annotated sentences from the same dialect showed the most marked increase, raising average performance by 4%. This effect tapers off quickly, showing no significant improvement between 500 and 1,000 training sentences with both settings performing at about 92% accuracy.

Training on multi-dialectal data had a much less marked effect. Using just the training sets of the other dialects, amounting to 4,000 sentences, has the same effect on performance as 125 from the same dialect. Doubling the multi-dialectal training data to 8,000 by using both training and test sets from the other dialects had no notable effect. Ultimately however, the best performance at 93% is attained with a combination of training data.

6. Conclusion and Future Work

We presented the MADAR CODA Corpus, a collection of 10,000 sentences from five Arabic city dialects (Beirut, Cairo, Doha, Rabat and Tunis) represented in the Conventional Orthography for Dialectal Arabic (CODA) in parallel with their Raw original form. We presented results on a

bootstrapping technique we used to speed up the CODA annotation, as well as on the degree of similarity across the dialects before and after CODA annotation. As expected CODA reduced the overall vocabulary within dialect and increased the overlap across dialects. The corpus will be publicly available for research purposes from `http://resources.camel-lab.com/`.

Our immediate next steps are to use the corpus to develop and benchmark systems for automatic multi-dialectal CODA annotation. We are also working on extending the size of the CODA corpus by adding more sentences from the MADAR corpus dataset, and diversifying with other data sets from Twitter (Bouamor et al., 2019) and speech recognition efforts (Ali et al., 2019).

Acknowledgments

This effort has been supported in part by the Multi-Arabic Dialect Applications and Resources (MADAR) project (grant NPRP 7-290-1-047 from the Qatar National Research Fund – a member of Qatar Foundation). All statements made herein are solely the responsibility of the authors.

7. Bibliographical References

Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Abu-Melhim, A.-R. (1991). Code-switching and linguistic accommodation in Arabic. In *Proceedings of the Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250.

Al-Badrashiny, M. and Diab, M. (2016). LILI: A simple language independent approach for language identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized Dialectal Arabic. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 30–38, Ann Arbor, Michigan.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and San'ani Yemeni Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Ali, A., Mubarak, H., and Vogel, S. (2014). Advances in dialectal Arabic speech recognition: A study using twitter to improve egyptian asr. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Ali, A., Khalifa, S., and Habash, N. (2019). Towards Variability Resistant Dialectal Speech Evaluation. In *Proc. Interspeech 2019*, pages 336–340.

Ali, A. M. A. M. (2018). *Multi-dialect Arabic broadcast speech recognition*. Ph.D. thesis, The University of Edinburgh.

Attia, M., Pecina, P., Samih, Y., Shaalan, K., and Van Genabith, J. (2016). Arabic spelling error detection and correction. *Natural Language Engineering*, 22(5):751–773.

Bassiouney, R. (2009). *Arabic Sociolinguistics: Topics in Diglossia, Gender, Identity, and Politics*. Georgetown University Press.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR Shared Task on Arabic Fine-Grained Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP19)*, Florence, Italy.

Cotterell, R. and Callison-Burch, C. (2014). A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 241–245, Reykjavik, Iceland.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3782–3789, Reykjavik, Iceland.

Erdmann, A., Habash, N., Taji, D., and Bouamor, H. (2017). Low Resourced Machine Translation via Morpho-syntactic Modeling: The Case of Dialectal Arabic. In *Proceedings of the Machine Translation Summit (MT Summit)*.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing spontaneous orthography. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 585–595, Atlanta, Georgia.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

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

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

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

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a Corpus for Palestinian Arabic: A Preliminary Study. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 18–27, Doha, Qatar.

Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., and Oflazer, K. (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.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A Large Scale Corpus of Gulf Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Kaabi, M. A. (2018). A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

McNeil, K. and Faiza, M. (2011). Tunisian Arabic Corpus : Creating a Written Corpus of an "Unwritten" Language. In *Proceedings of the Workshop on Arabic Corpus Linguistics (WACL)*.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 39–47, Doha, Qatar.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Rozovskaya, A., Bouamor, H., Habash, N., Zaghouani, W., Obeid, O., and Mohit, B. (2015). The second QALB shared task on automatic text correction for arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 26–35, Beijing, China.

Saadane, H. and Habash, N. (2015). A Conventional Orthography for Algerian Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, page 69, Beijing, China.

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014). YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1246–1251, Reykjavik, Iceland.

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

Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Watson, D., Zalmout, N., and Habash, N. (2018). Utilizing character and word embeddings for text normalization with sequence-to-sequence models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 837–843.

Watson, J. C. (2007). *The Phonology and Morphology of Arabic*. Oxford University Press.

Zaghouani, W. and Charfi, A. (2018). ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

Marcos Zampieri, et al., editors. (2019). *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, TOBEFILLED-Ann Arbor, Michigan, June. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine Translation of Arabic Dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 49–59, Montréal, Canada.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., and Habash, N. (2014). A conventional orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.