# Exploiting Dialect Identification in Automatic Dialectal Text Normalization

**Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor[†], Nizar Habash**
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
[†]Carnegie Mellon University in Qatar
{alhafni,sa5793,fan6236,za2051,fadhl.eryani,nizar.habash}@nyu.edu
hbouamor@cmu.edu

## Abstract

Dialectal Arabic is the primary spoken language used by native Arabic speakers in daily communication. The rise of social media platforms has notably expanded its use as a written language. However, Arabic dialects do not have standard orthographies. This, combined with the inherent noise in user-generated content on social media, presents a major challenge to NLP applications dealing with Dialectal Arabic. In this paper, we explore and report on the task of CODAfication, which aims to normalize Dialectal Arabic into the Conventional Orthography for Dialectal Arabic (CODA). We work with a unique parallel corpus of multiple Arabic dialects focusing on five major city dialects. We benchmark newly developed pretrained sequence-to-sequence models on the task of CODAfication. We further show that using dialect identification information improves the performance across all dialects. We make our code, data, and pretrained models publicly available.[1]

## 1 Introduction

Arabic exhibits a diglossic (Ferguson, 1959) linguistic situation where a non-standard variety, Dialectal Arabic (DA), coexists with Modern Standard Arabic (MSA), the standard form of the language. Complicating matters, there are multiple DA varieties, each differing from both other dialects and MSA in phonology, morphology, and lexicon. Arabic dialects are typically classified regionally, e.g., Egyptian, North African, Levantine, and Gulf. These dialects are the true native languages historically connected to Classical Arabic and other regional languages. While Arabic dialects are primarily spoken, they are increasingly used in written form on social media. Since Arabic dialects lack standard orthographies (Habash et al., 2018), DA text tends to be highly varied and noisy.

This high degree of noise poses major challenges for NLP systems as it increases the degree of sparsity in the data. Such noise can be handled using modeling techniques that normalize DA if it is used as an input to the system, e.g., in machine translation from dialects to other languages. However, challenges arise when the dialect itself is the desired output, for example, in automatic speech recognition systems (Ali et al., 2019; Sahyoun and Shehata, 2023). Consequently, evaluating and optimizing these systems can become problematic.

To mitigate the lack of orthographic standards for DA, several efforts in Arabic NLP introduced a common convention for DA spelling, named Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2018). However, the majority of approaches involving CODA consider it a side task to efforts like morphological disambiguation, diacritization, and lemmatization, as opposed to being the main target task.

Our contributions in this paper are as follows:

- We explore and report on the task of CODAfication (Eskander et al., 2013), normalizing DA text into the CODA convention. We work with a unique parallel corpus of multiple Arabic dialects (Eryani et al., 2020), focusing on five cities: Beirut, Cairo, Doha, Rabat, and Tunis.
- We benchmark newly developed pretrained sequence-to-sequence (Seq2Seq) models on the task of CODAfication.
- We demonstrate that using dialect identification information improves the performance across all dialects.

Next, we discuss some related work (§2) and then give a background on Arabic linguistic facts, CODA, and the data we use to train and test our models (§3). We describe our approach in §4 and present our experimental setup and results in §5.

[1] https://github.com/CAMeL-Lab/codafication



*Proceedings of The Second Arabic Natural Language Processing Conference*, pages 42–54

## 2 Related Work

### 2.1 Dialectal Arabic Text Normalization

DA NLP research has been receiving a considerable amount of attention, mainly due to the availability of monolingual and multilingual DA corpora (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; Bouamor et al., 2019). While MSA has well-defined orthographic standards, none of the Arabic dialects do today. As a result, almost all DA corpora were created without following any spelling conventions or standards, which are necessary for building robust DA NLP applications, e.g., machine translation (Erdmann et al., 2017). To mitigate this problem, several efforts have been introduced to standardize and develop orthographic conventions for Arabic dialects. Habash et al. (2012a) introduced the Conventional Orthography for Dialectal Arabic (CODA), the very first attempt to create guidelines and spelling conventions for Egyptian Arabic orthography. The convenience CODA offered by providing a standardized orthography led to the creation of many CODA extensions covering various dialects including Tunisian, Algerian, Palestinian, Moroccan, Yemeni, and Gulf Arabic (Zribi et al., 2014; Saadane and Habash, 2015; Jarrar et al., 2016; Turki et al., 2016; Khalifa et al., 2018). Each of these extensions tended to curate its own list of exceptional spellings for closed class words. Habash et al. (2018) introduced a unified set of guidelines for Arabic Dialect orthography – dubbed CODA* (CODA Star).

CODA has been used in the creation of a number of resources for DA NLP (Habash et al., 2012b; Eskander et al., 2013; Maamouri et al., 2014; Diab et al., 2014; Pasha et al., 2014b; Jarrar et al., 2016; Khalifa et al., 2018; Eryani et al., 2020). Most relevant to this paper is the work of Eryani et al. (2020) who extended a portion of the MADAR Corpus (Bouamor et al., 2018) to create the MADAR CODA Corpus, a collection of 10,000 sentences from five Arabic city dialects (Beirut, Cairo, Doha, Rabat, and Tunis) represented in the CODA standard in parallel with their original raw form. We use this corpus to train and test our models.

In terms of modeling approaches to CODAfication, the first work was proposed by Eskander et al. (2013) where they introduced CODAFY, a feature-based machine learning classifier to normalize Egyptian Arabic into CODA. Al-Badrashiny et al. (2014) and Shazal et al. (2020) targeted CODA output for dialectal Arabizi (Romanized Arabic) input. Most other approaches attempted to normalize DA texts into CODA as part of morphological analysis and disambiguation (Pasha et al., 2014a; Zalmout et al., 2018; Khalifa et al., 2020; Zalmout and Habash, 2020; Obeid et al., 2022). Our work is most similar to the one of Eskander et al. (2013) where we consider the task of CODAfication as a standalone text normalization task.

There has been some work on normalizing DA into MSA (Shaalan et al., 2007; Salloum and Habash, 2011, 2012; Alnajjar and Hämäläinen, 2024). While all this work is similar to ours in that dialectal input is processed, our output is still dialectal and not in MSA. Moreover, our proposed work has some similarities to grammatical error correction (GEC) for MSA (Zaghouani et al., 2014, 2015; Mohit et al., 2014; Rozovskaya et al., 2015; Watson et al., 2018; Habash and Palfreyman, 2022; Kwon et al., 2023; Alhafni et al., 2023). However, our task is different from GEC for MSA since GEC assumes a standard orthography that the writer is also assumed to aim for.

### 2.2 Dialect Identification

Dialect Identification (DID) is the task of identifying the dialect of a given speech or text fragment (Etman and Beex, 2015). Since informal conversations in real-world and online settings are typically conducted in DA, there has been a growing interest in developing and scaling automatic Arabic DID systems. This can be observed in the organization of multiple shared tasks (Bouamor et al., 2019; Abdul-Mageed et al., 2021, 2022, 2023, 2024) and the existence of various datasets and tools (Zaidan and Callison-Burch, 2011; Bouamor et al., 2014; Salama et al., 2014; Alsarsour et al., 2018; Abu Kwaik et al., 2018; Zaghouani and Charfi, 2018; Salameh et al., 2018; Bouamor et al., 2019; Abdelali et al., 2021; Baimukan et al., 2022). Besides its obvious use for profiling (Rangel et al., 2019), DA identification has already proved to be helpful for system selection in NLP tasks such as machine translation (Salloum et al., 2014), and morphological tagging (Obeid et al., 2022). In our work, we explore using text DID at the sentence-level in aiding CODAfication. For this, we use the CAMeL Tools (Obeid et al., 2020) DID implementation of Salameh et al. (2018).

| Dialect | Raw | CODA |
|---|---|---|
| Beirut | إزا بتريد ، تنان همبرغر و تنان أهوة . بدي آخدون معي .<br>*ĂzA btryd ، tnAn hmbrγr wtnAn Âhwħ . bdy Āxdwn mçy .* | اذا بتريد ، اثنين همبرغر واثنين قهوة . بدي آخذهن معي .<br>*AðA btryd ، Aθnyn hmbrγr wAθnyn qhwħ . bdy Āxðhn mçy .* |
| Cairo | اتنين هامبورجر و اتنين قهوة ، لوسمحت . عايزهم تيك اواي .<br>*Atnyn hAmbwrjr wAtnyn qhwħ ، lw smHt . çAyzhm tyk AwAy .* | اثنين هامبرجر واثنين قهوة ، لو سمحت . عايزهم تيك اوي .<br>*Aθnyn hAmbrjr wAθnyn qhwħ ، lw smHt . çAyzhm tyk Awy .* |
| Doha | اثنين همبرقر واثنين قهوة ، لو سمحت . باخذهم تيك اوي .<br>*Aθnyn hmbrqr wAθnyn qhwħ ، lw smHt . bAxðhm tyk Awy .* | اثنين همبرجر واثنين قهوة ، لو سمحت . بآخذهم تيك اوي .<br>*Aθnyn hmbrjr wAθnyn qhwħ ، lw smHt . bĀxðhm tyk Awy .* |
| Rabat | جوج هامبورغر و جوج قهيوات ، عافاك . غادي نديهوم معايا .<br>*jwj hAmbwrγr wjwj qhywAt ، çAfAk . γAdy ndyhwm mçAyA .* | جوج هامبورغر وجوج قهيوات ، عافاك . غادي نديهم معاي .<br>*jwj hAmbwrγr wjwj qhywAt ، çAfAk . γAdy ndyhm mçAy .* |
| Tunis | زوز همبرغر وزوز قهاوي ، يعيشك . نحب نهزهم معايا .<br>*zwz hmbrγr wzwz qhAwy ، yçyšk . nHb nhzhm mçAyA .* | زوز همبرغر وزوز قهاوي ، يعيشك . نحب نهزهم معاي .<br>*zwz hmbrγr wzwz qhAwy ، yçyšk . nHb nhzhm mçAy .* |

Table 1: An example sentence from the MADAR CODA Corpus in its raw and CODA parallel forms across five city dialects. The DA sentences are provided along with their transliterations in the HSB scheme (Habash et al., 2007). The sentence in the table can be translated as *"We would like two hamburgers and two coffees. To go, please."*

## 3 Background

### 3.1 Arabic Linguistic Facts

Arabic encompasses a wide range of dialectal varieties, with Modern Standard Arabic (MSA) serving as the common language of culture, media, and education across the Arab world. However, MSA is not the native language of any Arabic speaker, as dialectal Arabic dominates daily conversations. When native speakers write or speak (e.g., TV shows) in MSA, there is frequent code-mixing with the dialects in terms of phonological, morphological, and lexical choices (Abu-Melhim, 1991; Habash et al., 2008; Bassiouney, 2009). While Arabic dialects are typically classified regionally, e.g., Egyptian, North African, Levantine, and Gulf (Habash, 2010), hierarchical labels have been proposed to include countries, provinces, and cities (Baimukan et al., 2022). In this work, we focus on five city dialects: Beirut, Cairo, Doha, Rabat, and Tunis.

Despite their similarities, DA and MSA have many differences that prevent MSA tools from being effectively utilized for dialectal text. 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, none of the Arabic dialects do today. When Arabic speakers write in DA, they typically write in a way that reflects the phonology or the etymology of the words. Therefore, apart from unintentional typographical errors, no spelling of a dialectal word can be deemed truly "incorrect". This phenomenon is referred to as *spontaneous orthography* (Eskander et al., 2013; Eryani et al., 2020). For instance, the word for 'small [feminine singular]' in the Beirut dialect, /zʁiːri/, can be written in a range of spontaneous Arabic spellings, some of which highlighting its phonology and others its etymological connections to MSA صغيرة *Sγyrħ* /sˤaʁiːra[t]/. These include: زغيري *zγyry*, زغيره *zγyrh*, زغيرة *zγyrħ*, صغيري *Sγyry*, صغيره *Sγyrh*, and صغيرة *Sγyrħ*.

### 3.2 CODA

CODA*[2] (Habash et al., 2018) consolidates and standardizes several prior dialect-specific CODA conventions (Habash et al., 2012a; Saadane and Habash, 2015; Turki et al., 2016; Khalifa et al., 2016; Jarrar et al., 2016). CODA*, henceforth CODA, is an internally consistent and coherent convention for writing all DA varieties using the Arabic script aiming to balance dialectal uniqueness with MSA-DA similarities. CODA ensures consistency by controlling the natural spelling tendencies in spontaneous orthography that arise from writers considering etymological or phonological references of words. And while it is created for computational purposes, it is designed to be easily learnable and readable.

In the example mentioned above, the Beirut dialect word /zʁiːri/ 'small [feminine singular]' is written in a form reflective of MSA etymology: صغيرة *Sγyrħ*. Other examples of CODA from the MADAR CODA Corpus (Eryani et al., 2020) appear in Table 1. Note that foreign words pose a particular challenge to CODA due to the ambiguous phonological signals in the Arabic raw text. Conse-

[2]Pronounced *CODA Star*, as in, *for any dialect*.

| BEI | | | CAI | | | DOH | | | RAB | | | TUN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAW | CODA | FREQ | RAW | CODA | FREQ | RAW | CODA | FREQ | RAW | CODA | FREQ | RAW | CODA | FREQ |
| <SPC> | | 863 | <SPC> | | 1166 | | A ا | 150 | | A ا | 548 | | A ا | 458 |
| Â أ | A ا | 409 | Â أ | A ا | 608 | Â أ | A ا | 124 | A ا | | 352 | Â أ | A ا | 288 |
| Ă إ | A ا | 405 | h ه | ħ ة | 323 | | h ه | 82 | <SPC> | | 324 | Ă إ | A ا | 189 |
| | A ا | 324 | t ت | θ ث | 257 | ð ذ | Að اذ | 62 | Â أ | A ا | 256 | <SPC> | | 175 |
| t ت | θ ث | 294 | Ă إ | A ا | 146 | j ج | tš تش | 33 | t ت | θ ث | 190 | | <SPC> | 148 |
| | h ه | 173 | | <SPC> | 142 | j ج | k ك | 31 | | <SPC> | 168 | A ا | ħ ة | 115 |
| w و | h ه | 138 | d د | ð ذ | 95 | | <SPC>ا<br><SPC>A | 28 | A ا | ħ ة | 160 | A ا | | 109 |
| Â أ | q ق | 129 | ý ى | y ي | 80 | A ا | Ā آ | 25 | Ă إ | A ا | 84 | | l ل | 100 |
| d د | ð ذ | 119 | | A ا | 73 | Ă إ | A ا | 23 | d د | ð ذ | 68 | w و | h ه | 97 |
| | <SPC> | 106 | A ا | | 68 | y ي | j ج | 20 | | l ل | 67 | | n ن | 85 |

Table 2: The top 10 character edit transformations from raw to CODA in the entire MADAR CODA dataset across the five dialects. <SPC> indicates an explicit white space; whereas an empty cell indicates a *null* string.

quently, Eryani et al. (2020) adopted a minimalistic strategy for CODAfying these words, resulting in some plausible but inconsistent variants. For example, the word for 'hamburger' in Table 1 appears as both همبرغر *hmbrγr* and همبرجر *hmbrjr*.

### 3.3 MADAR CODA Corpus

We use the manually annotated MADAR CODA Corpus (Eryani et al., 2020), a collection of 10,000 sentences from five Arabic city dialects (Beirut, Cairo, Doha, Rabat, and Tunis) represented in the CODA standard in parallel with their original raw form. The sentences come from the Multi-Arabic Dialect Applications and Resources (MADAR) Project (Bouamor et al., 2018) and are in parallel across the cities (2,000 sentences from each city).

The corpus is originally split into train and test, with each split consisting of 5,000 parallel sentences (1,000 per dialect). In our setup, we combine the original train and test splits and then divide the data randomly into separate training (Train), development (Dev), and testing (Test) sets. We use a 70/15/15 split, resulting in 1400, 300, and 300 sentences, respectively, per dialect. In total, we end up with 7,000 sentences for Train, 1,500 for Dev, and 1,500 for Test. Table 1 shows an example of a sentence from the corpus in its raw and CODA parallel forms across the five city dialects.

Table 2 presents the top 10 character-level edit changes from raw text to CODA in the five city dialects. It is noteworthy that while there are many shared transformations, they appear with different distributions. This suggests that a model making use of DID could learn dialect-specific preferences. At the same time, the shared phenomena can aid in learning dialect-independent general patterns.

## 4 Approach

We frame the CODAfication task as a controlled text generation problem. Formally, given a dialectal input sentence $X$ and its dialect $D$, the goal is to generate the CODAfied sentence $Y$ according to $P(Y|X, D)$. One way to condition text generation models on the desired dialect, $D$, is to represent it as a special "control" token appended to the input sequence $[D; X]$, which acts as a side constraint (Sennrich et al., 2016a). In Seq2Seq models, this allows the encoder to learn a representation for this token as any other token in its vocabulary, and the decoder attends to this representation to guide the generation of the output sequence.

This simple strategy has been used in various controlled text generation tasks such as machine translation (Sennrich et al., 2016b; Sennrich and Haddow, 2016; Johnson et al., 2016; Agrawal and Carpuat, 2019), style transfer (Niu et al., 2017, 2018), text simplification (Yanamoto et al., 2022; Agrawal and Carpuat, 2023), and Arabic gender rewriting (Alhafni et al., 2022).

We experiment with two recently developed pretrained Arabic Transformer-based Seq2Seq models: AraBART (Kamal Eddine et al., 2022), which was pretrained on 24GB of MSA data primarily from the news domain, and AraT5-v2 (Nagoudi et al., 2022; Elmadany et al., 2023), which was pretrained on a larger dataset of 250GB covering MSA, DA, and Classical Arabic (CA) data.

We explore using four different control tokens to pass the dialect information to the models. Table 3 presents the control tokens we considered in our experiments:

- **City**: The name of the city where the Arabic dialect is spoken.

| Dialect | City | MSA Phrase | DA Phrase | Digit |
|---|---|---|---|---|
| Beirut | بيروت<br>*byrwt* | في بيروت نقول<br>*fy byrwt nqwl* | في بيروت منقول<br>*fy byrwt mnqwl* | 1 |
| Cairo | القاهرة<br>*AlqAhrħ* | في القاهرة نقول<br>*fy AlqAhrħ nqwl* | في القاهرة بنقول<br>*fy AlqAhrħ bnqwl* | 2 |
| Doha | الدوحة<br>*AldwHħ* | في الدوحة نقول<br>*fy AldwHħ nqwl* | في الدوحة نقول<br>*fy AldwHħ nqwl* | 3 |
| Rabat | الرباط<br>*AlrbAT* | في الرباط نقول<br>*fy AlrbAT nqwl* | في الرباط كنقولو<br>*fy AlrbAT knqwlw* | 4 |
| Tunis | تونس<br>*twns* | في تونس نقول<br>*fy twns nqwl* | في تونس نقولو<br>*fy twns nqwlw* | 5 |

Table 3: The four different types of control tokens we use in our experiments.

- **MSA Phrase**: An MSA phrase that follows the template في <*city*> نقول 'in <*city*> we say', where <*city*> represents one of the five cities whose dialects we are modeling.
- **DA Phrase**: A DA phrase that follows the template <*we-say*> <*city*> في 'in <*city*> we say', where <*city*> represents one of the five dialects we are modeling, and <*we-say*> represents a spontaneous orthography of the dialectal version of the phrase 'we say'.
- **Digit**: An ad hoc unique numerical value for each dialect.

During training, we use the gold dialect for each sentence to induce its control tokens. To obtain the dialect during inference, we use the DID system that is available in CAMeL Tools (Obeid et al., 2020). The system is an implementation of Salameh et al. (2018)'s best-performing model on the MADAR shared task on DID (Bouamor et al., 2019). The system models DID for the five city dialects and MSA. We fine-tune the Seq2Seq models on a single GPU for 10 epochs, a batch size of 16, and a maximum sequence length of 200 using Hugging Face's Transformers (Wolf et al., 2019). We use learning rates of 5e-5 and 1e-4, for AraBART and AraT5, respectively. During inference, we use beam search with a beam width of 5.

## 5 Experimental Setup and Results

### 5.1 Metrics

We use the MaxMatch ($M^2$) scorer (Dahlmeier and Ng, 2012), which is predominantly used to evaluate grammatical error correction systems. The $M^2$ scorer assesses the edits made by the system against the 'gold standard' edits in the target CODA, calculating precision (P), recall (R), $F_1$, and $F_{0.5}$ scores. $F_{0.5}$ weighs precision twice as much as recall, to prioritize the accuracy of edits relative to all edits made by the system. To obtain the gold edits, we use the alignment algorithm that was proposed by Alhafni et al. (2023). We also use their optimized version of the $M^2$ scorer that deals with the extreme running times of the original release in cases where the generated outputs differ significantly from the input.

Moreover, and to be consistent with previous work, we report the Word Error Rate (WER). However, we believe that WER is not a suitable metric for the task of CODAfication due to the high similarity between the input and output sentences.

### 5.2 Models: Baselines and Systems

**Do Nothing** Our first baseline simply copies the input sentences to the output. This baseline highlights the level of similarity between the inputs and outputs.

**Maximum Likelihood Estimation (MLE)** For the second baseline, we build a simple word-level lookup model to map input words to their CODAfied versions. We first obtain word-level alignments over all the training data from all the dialects (**Joint**) by using the algorithm developed by Alhafni et al. (2023). We then exploit the alignments to implement the lookup model as a bigram maximum likelihood estimator: given an input word with its bigram surrounding context ($w_i$, $w_{i-1}$), and a CODAfied target word ($y_i$), the model is built by computing $P(y_i|w_i, w_{i-1})$ over the training examples. During inference, we generate all possible alternatives for the given input word

| Model | Training | Control Token | P | R | F1 | F0.5 | WER |
|---|---|---|---|---|---|---|---|
| **Do Nothing** | - | - | 100 | 0 | 0 | 0 | 0.2677 |
| **MLE** | **Joint** | - | 66.81 | 44.62 | 53.51 | 60.77 | 0.1456 |
| **AraT5** | **Joint** | - | 86.76 | 77.44 | 81.83 | 84.72 | 0.0620 |
| | | **City** | **87.58** | **79.34** | **83.25** | **85.80** | **0.0566** |
| | | | 87.52 | 79.34 | 83.23 | 85.75 | 0.0566 |
| | | **MSA Phrase** | 87.43 | 79.06 | 83.03 | 85.62 | 0.0573 |
| | | | 87.43 | 79.06 | 83.03 | 85.62 | 0.0572 |
| | | **DA Phrase** | 87.25 | 78.56 | 82.68 | 85.36 | 0.0601 |
| | | | 87.26 | 78.61 | 82.71 | 85.38 | 0.0602 |
| | | **Digit** | 87.37 | 79.00 | 82.98 | 85.56 | 0.0588 |
| | | | 87.37 | 79.00 | 82.98 | 85.56 | 0.0588 |
| | **Ensemble** | - | 85.65 | 72.84 | 78.73 | 82.74 | 0.0739 |
| | | | 85.52 | 73.07 | 78.80 | 82.70 | 0.0730 |
| **AraBART** | **Joint** | - | 85.35 | 74.36 | 79.47 | 82.90 | 0.0728 |
| | | **City** | 85.69 | 74.41 | 79.65 | 83.17 | 0.0715 |
| | | | 85.64 | 74.47 | 79.66 | 83.15 | 0.0715 |
| | | **MSA Phrase** | 85.48 | 74.47 | 79.59 | 83.02 | 0.0716 |
| | | | 85.48 | 74.47 | 79.59 | 83.02 | 0.0716 |
| | | **DA Phrase** | 85.00 | 74.58 | 79.45 | 82.69 | 0.0721 |
| | | | 85.00 | 74.58 | 79.45 | 82.69 | 0.0721 |
| | | **Digit** | 86.11 | 73.96 | 79.58 | 83.38 | 0.0732 |
| | | | 86.11 | 73.96 | 79.58 | 83.38 | 0.0731 |
| | **Ensemble** | - | 84.59 | 67.92 | 75.34 | 80.63 | 0.0843 |
| | | | 84.40 | 68.48 | 75.61 | 80.65 | 0.0829 |

Table 4: Results of number of systems on the Dev set. Results in grey indicate using gold DID labels (i.e., Oracle). Bolding indicates the best results. Best results in the oracle setup are underlined.

($w_i$). If the bigram context ($w_i$, $w_{i-1}$) was not observed in the training data, we backoff to a unigram context. If the input word was not observed during training, we pass it to the output as it is.

**Seq2Seq** We train both AraBART and AraT5 on all the dialects' training data jointly with and without using DID information. We refer to this modeling setup as **Joint**. Moreover, to examine the effect of the joint dialectal training, we train five separate models, one for each dialect. During inference, we combine the separate models in an ensemble setup where we use the DID predictions for each sentence to select the appropriate model. We refer to this setup as **Ensemble**.

### 5.3 Results

**Overall Results** Table 4 presents the results on the Dev set. Among the baselines, both AraBART and AraT5 demonstrate superior performance compared to the MLE model. In terms of training setups, **Joint** training outperforms **Ensemble** models for both AraBART and AraT5, with AraT5 being the better performer achieving 84.72 $F_{0.5}$.

When we train the AraBART **Joint** variants with DID control tokens, the performance increases compared to the AraBART **Joint** baseline, except when training with the **DA Phrase** DID control token. All the AraT5 **Joint** variants benefit from training with DID control tokens compared to the AraT5 baseline, with the **City** control token being the best performer with 85.80 $F_{0.5}$ (1.08 increase over the AraT5 baseline and statistically significant at $p < 0.05$).[3] It is noteworthy that the AraT5 variants perform better compared to their AraBART counterparts across all experiments. We suspect this is due to the fact the data used to pretrain AraT5 consisted of a mix of MSA, DA, and CA compared to only MSA in the case of AraBART's pretraining.

Since AraT5 performed better than AraBART across all experiments, we present the results on

[3]Statistical significance was done using a two-sided approximate randomization test.

| Model | Training | Control Token | P | R | F1 | F0.5 | WER |
|---|---|---|---|---|---|---|---|
| AraT5 | Joint | - | 87.26 | 77.98 | 82.36 | 85.23 | 0.0622 |
| | | City | 87.99 | 78.25 | 82.83 | 85.85 | 0.0601 |
| | | MSA Phrase | 88.17 | 78.85 | 83.25 | 86.13 | 0.0583 |
| | | DA Phrase | **88.35** | **78.95** | **83.39** | **86.29** | **0.0573** |
| | | Digit | 87.67 | 78.25 | 82.69 | 85.61 | 0.0610 |

Table 5: Results of the AraT5 variants on the Test set.

| | AraT5 (Baseline) | | | | | AraT5 + City | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dialect | P | R | F1 | F0.5 | WER | P | R | F1 | F0.5 | WER |
| Beirut | 86.07 | 79.70 | 82.76 | 84.71 | 0.0829 | **89.30** | **82.35** | **85.69** | **87.82** | **0.0673** |
| Cairo | **89.52** | **85.42** | **87.42** | **88.67** | **0.0582** | 89.13 | **85.42** | 87.23 | 88.36 | 0.0588 |
| Doha | 83.52 | 67.86 | 74.88 | 79.83 | 0.0302 | **85.26** | **72.32** | **78.26** | **82.32** | **0.0277** |
| Rabat | 85.21 | 72.67 | 78.44 | 82.37 | 0.0557 | **86.35** | **75.98** | **80.83** | **84.05** | **0.0493** |
| Tunis | **86.08** | 70.36 | **77.43** | **82.40** | 0.0821 | 84.15 | **71.56** | 77.35 | 81.29 | **0.0792** |

Table 6: Dialect-specific results of the best system (AratT5 + City) against the baseline (AraT5) on the Dev set.

| | AraT5 (Baseline) | | | | | AraT5 + DA Phrase | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dialect | P | R | F1 | F0.5 | WER | P | R | F1 | F0.5 | WER |
| Beirut | 85.57 | 78.54 | 81.91 | 84.07 | 0.0847 | **87.02** | **80.04** | **83.38** | **85.53** | **0.0760** |
| Cairo | **89.65** | 83.43 | 86.43 | 88.33 | 0.0609 | 89.57 | **84.39** | **86.90** | **88.48** | **0.0592** |
| Doha | 85.86 | 70.83 | 77.63 | 82.36 | 0.0295 | **89.80** | **73.33** | **80.73** | **85.94** | **0.0245** |
| Rabat | 87.11 | **73.87** | **79.94** | 84.09 | 0.0685 | **87.34** | 73.60 | 79.88 | **84.20** | **0.0674** |
| Tunis | 86.74 | 75.58 | 80.78 | 84.25 | 0.0661 | **89.23** | **76.57** | **82.42** | **86.37** | **0.0577** |

Table 7: Dialect-specific results of the best system (AratT5 + DA Phrase) against the baseline (AraT5) on the Test set.

the Test set using AraT5 and its variants in Table 5. Training AraT5 with the **DA Phrase** control token yields the best performance on the Test set with 86.29 $F_{0.5}$ (1.06 increase over the AraT5 baseline and statistically significant at $p < 0.05$).

**DID Efficacy** We estimate an oracle upper bound by using gold DID labels during inference on the Dev set (Table 4). We do not notice significant improvements across all variants compared to the models that use predicted DID labels. In some cases, using gold DID labels results in identical performance to models using predicted labels. This can be attributed to the robustness of our CODAfication models and the reliability of the DID system we are using, which achieves a high accuracy of 92.1% on the Dev set.

Most of the prediction errors made by the DID system occur in sentences lacking distinctive cues that would allow clear assignment to a specific dialect. Therefore, these errors cannot be considered true errors, but rather stem from the MADAR dataset's limitation of not having multi-dialectal labels. This is consistent with the findings of Keleg and Magdy (2023) where they manually analyzed the errors of a single-label DID system and found that ∼66% of the errors are not true errors and could be resolved with multi-dialect labels.

**Dialect-Specific Results** We present the dialect-specific results on the Dev and Test sets in Tables 6 and 7, respectively. Our best system on the Dev set, AraT5 trained with the **City** DID control token, improves the results over the AraT5 baseline for all dialects (with the largest increase seen for Beirut at 3.11 $F_{0.5}$), except for Cairo and Tunis, where the performance drop is attributed to decreased precision rather than recall. This suggests that our best system may be making unnecessary extra rewrites. On the Test set, our best system, AraT5 trained with the **DA Phrase** DID control token, improves the results over the AraT5 baseline across all dialects, with the largest increase for Doha at 3.58 $F_{0.5}$.

| Category | % | Error | CODA |
|---|---|---|---|
| **Non-CODA** | 46% | تحدست<br>*tHdst* | تحدثت<br>*tHdθt* |
| **Hallucination** | 19% | ديقة.<br>*dyqħ.* | دقيقة.<br>*dqyqħ.* |
| **Valid** | 13% | هامبورجر<br>*hAmbwrjr* | هامبرجر<br>*hAmbrjr* |
| **Deletion** | 9% | اوصلة<br>*AwSlħ* | اوصل له<br>*AwSl lh* |
| **Related Hallucination** | 9% | شرف<br>*šrf* | الشرف<br>*Alšrf* |
| **Punctuation** | 4% | فاتتني"""""""""<br>*fAttny"""""""""* | فاتتني"""<br>*fAttny"""* |

Table 8: Distribution of errors in the Dev set with one example per error type.

### 5.4 Error Analysis

To gain insights into the errors present in our best performing system on the Dev set, we conducted an error analysis on a sample of 100 cases, which accounted for 21% of the total 471 erroneous instances in the generated output. We classified these errors into specific categories, with results and examples provided in Table 8:

- **Non-CODA**: These are cases characterized by having plausible spontaneous spelling but incorrect CODA. This is the largest group of errors.
- **Hallucination and Related Hallucination**: Hallucinations refer to word rewrites that are implausible under any circumstance as a CODA correction or non-CODA spelling. We distinguish cases that seem morphologically related to the input but are actually unrelated forms. We observe that 2/3 of the cases were largely unrelated to the reference.
- **Valid**: This category encompasses valid alternative spellings, particularly those associated with proper nouns and foreign words.
- **Deletion**: Deletions refer to omitted words. 55.6% of these are non-CODA spellings, e.g., a missed split (Table 8 example), while the rest are divided between gold errors and hallucinations.
- **Punctuation**: Punctuation generation errors.

The error analysis highlights that CODA issues constitute a significant portion of the remaining errors, potentially accounting for half of the cases between non-CODA words and deletions. Hallucinations, whether minor or severe, make up nearly a third of the errors. These findings suggest the need for more training data and improved models to address these problems. The presence of valid variants, which represent one-eighth of the errors, indicates the need to adopt a multi-reference approach for text normalization evaluation.

## 6 Conclusion and Future Work

We explored and reported on the task of CODAfication, i.e., normalizing Dialectal Arabic into the Conventional Orthography for Dialectal Arabic (CODA). We benchmarked newly developed pretrained Seq2Seq models on the task of CODAfication. We further showed, for the first time to our knowledge, that using dialect identification information improves Arabic text normalization.

In future work, we plan to explore other modeling approaches, including multitask learning models for both DID and CODAficiation, as well as text editing models (Omelianchuk et al., 2020, *inter alia*). We also plan to extend our work to CODA data sets for other dialects (Jarrar et al., 2016; Khalifa et al., 2016), evaluate the added value of improved CODAfication on downstream NLP tasks, and develop models of CODA error type classification (Belkebir and Habash, 2021).

## Limitations

Although we benchmarked pretrained Seq2Seq models on the task of CODAfication and demonstrated the added improvements of using dialect identification information, we did not conduct experiments to showcase the added value of the task of CODAfication on downstream NLP tasks such as sentiment analysis and machine translation. The efficacy of CODAfication in enhancing these downstream applications, particularly with newer models, remains an area for future exploration.

Our work is based on a unique curated parallel corpus encompassing multiple Arabic dialects from five cities. While this dataset provides valuable insights into CODAfication performance across diverse dialectal variations, it also introduces limitations in generalizing our findings to a broader spectrum of Arabic dialects beyond our specific dataset. Future research should aim to extend the evaluation of CODAfication models across a more extensive range of dialectal datasets to ensure robustness and applicability across diverse linguistic contexts.

## Acknowledgements

We acknowledge the support of the High Performance Computing Center at New York University Abu Dhabi. We thank the anonymous reviewers for their feedback and suggestions.

## References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

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

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic transliteration of Romanized dialectal Arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 30–38, Ann Arbor, Michigan. Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Ahmed Ali, Salam Khalifa, and Nizar Habash. 2019. Towards Variability Resistant Dialectal Speech Evaluation. In *Proc. Interspeech 2019*, pages 336–340.

Khalid Alnajjar and Mika Hämäläinen. 2024. Normalization of arabic dialects into modern standard arabic using. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.

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

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A Human Judgement Corpus and a Metric for Arabic MT Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar.

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

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

Ryan Cotterell and Chris Callison-Burch. 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.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Mona T Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 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.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low resourced machine translation via morpho-syntactic modeling: The case of dialectal Arabic. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 185–200, Nagoya Japan.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4130–4138, Marseille, France. European Language Resources Association.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 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.

Asma Etman and Louis Beex. 2015. Language and Dialect Identification: A Survey. In *Proceedings of the Intelligent Systems Conference (IntelliSys)*, London, UK.

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

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, 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, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the Workshop on HLT & NLP within the Arabic World*, Marrakech, Morocco.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. *On Arabic Transliteration*, pages 15–22. Springer Netherlands, Dordrecht.

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

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

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.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

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

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 Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Salam Khalifa, Nasser Zalmout, and Nizar Habash. 2020. Morphological analysis and disambiguation for Gulf Arabic: The interplay between resources and methods. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3895–3904, Marseille, France. European Language Resources Association.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

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).

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

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 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.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–326, Abu Dhabi, UAE. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014a. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014b. 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.

Francisco Rangel, Paolo Rosso, Anis Charfi, Wajdi Zaghouani, Bilal Ghanem, and Javier Sánchez-Junquera. 2019. On the author profiling and deception detection in Arabic shared task at FIRE. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 7–9.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 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.

Houda Saadane and Nizar Habash. 2015. A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, page 69, Beijing, China.

Abdulwahab Sahyoun and Shady Shehata. 2023. AraDiaWER: An explainable metric for dialectical Arabic ASR. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 64–73, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778, Baltimore, Maryland. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Wael Salloum and Nizar Habash. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 385–392, Mumbai, India.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, volume 1, pages 83–91.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

K. Shaalan, H.M.A. Bakr, and I. Ziedan. 2007. Transferring Egyptian colloquial dialect into modern standard Arabic. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 525–529.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.

Houcemeddine Turki, Emad Adel, Tariq Daouda, and Nassim Regragui. 2016. A Conventional Orthography for Maghrebi Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Daniel Watson, Nasser Zalmout, and Nizar Habash. 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, Brussels, Belgium. Association for Computational Linguistics.

Janet CE Watson. 2007. *The Phonology and Morphology of Arabic*. Oxford University Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 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.

Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider, and Kemal Oflazer. 2015. Correction annotation for non-native Arabic texts: Guidelines and corpus. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 129–139.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Omar F Zaidan and Chris Callison-Burch. 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.

Nasser Zalmout, Alexander Erdmann, and Nizar Habash. 2018. Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 953–964, New Orleans, Louisiana. Association for Computational Linguistics.

Nasser Zalmout and Nizar Habash. 2020. Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8297–8307, Online. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 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.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.