INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards Variability Resistant Dialectal Speech Evaluation

Ahmed Ali,[†] *Salam Khalifa,*[‡] *Nizar Habash*[‡]

[†]Qatar Computing Research Institute
Hamad Bin Khalifa University, Doha, Qatar
[‡]Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
amali@qf.org.qa, salamkhalifa@nyu.edu, nizar.habash@nyu.edu

Abstract

We study the problem of evaluating automatic speech recognition (ASR) systems that target dialectal speech input. A major challenge in this case is that the orthography of dialects is typically not standardized. From an ASR evaluation perspective, this means that there is no clear gold standard for the expected output, and several possible outputs could be considered correct according to different human annotators, which makes standard word error rate (WER) inadequate as an evaluation metric. Specifically targeting the case of Arabic dialects, which are also morphologically rich and complex, we propose a number of alternative WER-based metrics that vary in terms of text representation, including different degrees of morphological abstraction and spelling normalization. We evaluate the efficacy of these metrics by comparing their correlation with human judgments on a validation set of 1,000 utterances. Our results show that the use of morphological abstractions and spelling normalization produces systems with higher correlation with human judgment. We released the code and the datasets to the research community.

Index Terms: ASR, Dialects, Non-standard Orthography, Evaluation, Metrics

1. Introduction

Automatic Speech Recognition (ASR) has shown fast progress recently, thanks to advancements in deep learning. As a result, the best systems for English have achieved a single-digit word error rate (WER) for some conversational tasks [1]. However, this is different for dialectal ASR, for which the WER can easily go over 40% [2]. While the size of available data and the morphological complexity of the language affect the quality of the ASR system, we focus here on the challenges brought by lack of spelling standardization in dialects. In a standardized language such as English, we know that *enough* is a correct spelling, while *enuf* is not. However, for languages and dialects without official standards, the same confident sentiment about spelling cannot be expressed as there is no "correct" spelling, but rather a range of variations; at best, we would know what a preferred or a dominant spelling is [3, 4]. In this paper we exclusively target intra-dialectal variation and not inter-dialect variation, i.e., among different Arabic dialects.

Table 1 shows some examples of spelling variation in Dialectal Arabic (DA). We can see that clitics (pronouns and negations) can be written concatenated or separated from the verb, the definite article can undergo different spelling variations due to coarticulation with the following word, long vowels can become short, and thus be dropped as they are typically not written in Arabic, etc. While some variations can happen in standardized languages such as English, e.g., *healthcare* vs. *health care*, or *organize* vs. *organise*, this is much less common, and in ASR it is easily handled with simple rules, e.g., the Global Mapping file in SCLITE [5, 6]. However, there are far more acceptable spelling variants in DA; for instance, [7] used 11 million pairs extracted from a seven-billion words corpus of DA tweets.

Table 1: *Egyptian phrases with multiple spelling variants: shown in Arabic script and in Buckwalter transliteration [10].*

English Gloss	IPA	Spelling Variants			
'he was not'	/makánʃ/	ما كانش	ماكانش	ماكنش	مكنش
		mA kAn$	*mAkAn$*	*mAkn$*	*mkn$*
'I told him'	/ʔultɛ́lo/	قلت له	قولتله	قولت له	قلتله
		qlt lh	*qwltlh*	*qwlt lh*	*qltlh*
'by the morning'	/ʕassóbħ/	عالصبح	ع الصبح	علصبح	عصّبح
		EAlSbH	*E AlSbH*	*ElSbH*	*ESbH*

The above examples partially explain the high WER for dialects. While they suffer from the lack of training resources, the main problem is their informal status, which results in rarely regulated spelling. This makes training an ASR system for dialects much harder as there is no single gold standard towards which to optimize at training time. More importantly, it is hard to evaluate such a system and to measure progress as multiple possible text outputs for the same speech signal could be considered correct by different people. Thus, there is need for evaluation measures that allow for common spelling variations.

Previously, the problem was addressed using the multi-reference word error rate (WER_{mr}) [8], which is similar to the multi-reference BLEU score [9] used to evaluate Machine Translation (MT). However, obtaining multiple references is expensive. Moreover, it could take many human annotators to get good coverage of the possible orthographic variants of the transcription of a speech recording. In a further study the same problem was addressed using single reference, but in the process of comparing a hypothesis to a reference, the introduced dialectal word error rate (WER*d*) [7] makes use of spelling variants for words and phrases, which was mined from Twitter in an unsupervised fashion. The experiments with evaluating ASR output for Egyptian Arabic, and further manual analysis, show that the resulting WERd metric, is more adequate than WER for evaluating dialectal ASR.

In the last few years, a number of researchers working on Arabic NLP proposed a *Conventional Orthography for Dialectal Arabic* (CODA) as a systematic way to spell DA words for a number of dialects [11, 12, 4]. This system has been used in a number of datasets [13, 14] and tools for Arabic automatic processing, such as MADAMIRA [15]. In Table 1, the CODA form is the first from the left. In addition to the orthographic variability challenge, Arabic and its dialects are morphologically rich and complex languages often including hundreds of inflected

http://dx.doi.org/10.21437/Interspeech.2019-2692

forms for a single *lemma* (base or dictionary form) [16].

In this paper we explore evaluating using alternative representations that could reduce some of the variation in human transcriptions of dialectal speech. To that end we compare the use of CODA and automatically generated morphologically abstracted forms (tokenizations and lemmatization) with raw text and simple text normalization techniques. We pair varying representations with varying multi-reference evaluation techniques. We evaluate the efficacy of such approaches through correlation with human judgments on a validation set. Our contributions are: (i) a novel approach to reduce the variance in evaluating dialectal ASR; (ii) an evaluation of our approach which compares the correlation with processing WER_{avg} (average WER) and WER_{mr} (Multi-reference WER); and (iii) a release of the paper's data and code for further research.[1]

2. Arabic Dialect Spontaneous Orthography

2.1. Arabic and its dialects

Arabic is a collection of variants among which one particular variant has a special status as the formal written standard of the media, culture and education across the Arab World – Modern Standard Arabic (MSA). Although it is the national language, MSA is not the native language of any modern day Arabs. The other variants are informal spoken dialects that are the media of communication for daily life, and true native languages [16].

2.2. Writing without a standard orthography

Given that DA is the commonly spoken form of Arabic, it is naturally used in its written form throughout the web and specifically the different social media platforms. Since there are no published orthography standards for the dialects, people tend to write spontaneously, either representing the phonology of the word, referring to its MSA etymological cognate or a mix of both. [4] showed that there are at least 27 encountered ways to write the Egyptian Arabic word /mabiʔulha:ʃ/ 'he does not say it', including, مبيقولهاش *mbyqwlhA$* and مبؤلهاش *mb&lhA$*. DA text is also sometimes written using a non-standard romanization known as Arabizi [17, 18].

2.3. Relevant differences between EGY and MSA

Similar to many other Arabic dialects, the distance of Egyptian Arabic (EGY) from MSA varies among different dimensions. Phonologically, some of the important differences between EGY and MSA [16, 11] include: (i) the MSA sounds written as the letters ج *j* , ق *q* , ث *v* , ذ * and ظ *Z* (/ʤ/, /q/, /θ/, /ð/, and /ðˤ/) realize differently in EGY as /g/, /ʔ/, /t/ or /s/, /d/ or /z/, and /zˤ/, respectively; (ii) a change or a complete drop in short vowels; and (iii) predictable shortening of long vowels in certain word positions. Morphologically, EGY simplifies parts of the MSA paradigms, such as the loss of case markings for nominals and mood and voices for verbs. However, EGY introduces new clitics that add to the morphological complexity, such as the negation circum-clitic ش+ +ما *mA+ +$*. Finally, there are many lexical variations, resulting in very different distributions that affect how language models and embeddings can be used multi-dialectally [19], e.g., the **primary** sense of the word عربية *Erbyp* is 'female Arab' in MSA, but 'car' in EGY.

[1] https://github.com/qcri/ArabicSpeechTextProcessing

3. Datasets

3.1. The MGB-3 Arabic dataset

The MGB-3 Arabic data comprised 16 hours of Egyptian Arabic speech extracted from 80 YouTube videos distributed across seven genres: comedy, cooking, family/kids, fashion, drama, sports, and science talks (TEDx). In the 2017 *Speech Recognition Challenge in the Wild: Arabic MGB-3* [2], it was assumed that the MGB-3 data is not enough by itself to build robust speech recognition systems, but could be useful for adaptation, and for hyper-parameter tuning of models built using the MGB2 [20] data. Therefore, participants reused the MGB2 training data in that challenge, and considered the provided in-domain data as (supervised) adaptation data. The challenge goal was to build a standard speech-to-text transcription system and to provide time-stamped word recognition results. The system outputs we use in this paper were produced in this challenge. See [2] for more details.

3.2. The multi-transcription dataset

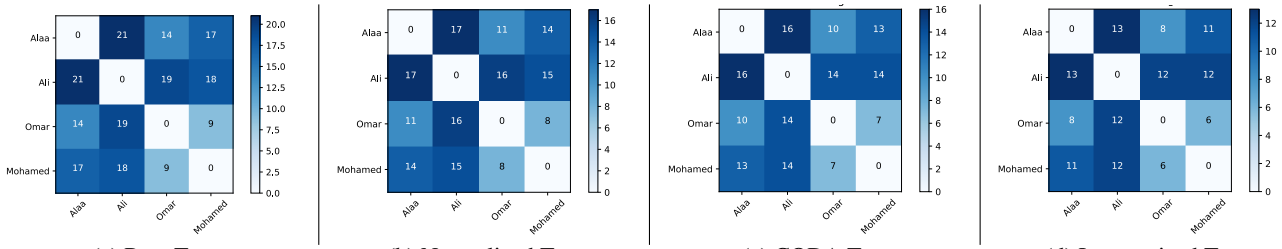
In contrast with the approach taken by the CODA designers, we decided to have four transcriptions, allowing transcribers to write the transcripts as they deemed correct. Figure 1 shows the inter-annotator disagreement on the 2,500 sentences in the test data. There is between 9-21% disagreement between the annotators for the raw test data. We can reduce the disagreement by normalizing the text using commonly done Alif/Ya/Ta-Marbuta normalization. This bring down the disagreement to between 8-17%. This is still rather high given that these are all in principle 'gold' references. We convert the text to CODA and lemmatize it automatically using the MADAMIRA system which supports both MSA and EGY [15]. The CODA text has reduced inconsistency in transcription with respect to the normalized text (7-16%). Of course, lemmatization reduces the inter annotation disagreement further (6-13%). From these results we will consider these four representations for evaluating dialectal speech transcription in addition to a couple of other contrastive representations known in the literature [21, 16].

3.3. The human judgments

Since it is easy in principle to make a WER go down with extensive text normalization (in the absurd extreme turning everything into a single symbol), comparing correlations of ranks with humans, even if the humans had some disagreement, is a good indicator of the value of a particular setup.

We created a human rank judgment validation set for 1,000 randomly selected MGB-3 sentences with multiple transcriptions, this is 2.1 hours. We evaluated the leading five ASR systems in the MGB-3: Aalto University (AALTO), NDSC-THUEE, Johns Hopkins University (JHU), MIT, and Brno University of Technology (BUT). In an attempt to validate our findings, we considered one of the reference transcriptions (Mohamed in Figure 1, henceforth *Human*) as system. For more details about the ASR systems and results, see [2].

We used four human judges, all of whom are educated native Egyptian speakers with extensive experience in linguistic annotation. We used Qualtrics for the ranking interface. For each sentence the assigned judge saw the unique instances of three human transcriptions from the multi-reference set discussed above. The fourth human transcription was included as a system (*Human*). The unique instances produced by the six above mentioned systems (total between 1 and 6) appear under the references in a different section with a different font size and



(a) Raw Text (b) Normalized Text (c) CODA Text (d) Lemmatized Text

Figure 1: *Inter annotation disagreement for the four transcribers across four representations.*

Table 2: *Scores of different metrics and representations (i) and their resulting rankings (ii) on six systems (five ASR and one Human) on our validation set. Red cells mark instances of disagreement between human rank and metric rank. The last three rows in (i) show the difference between* RAW$_{norm}$ *and* RAW, CODA *and* RAW, *and the relative reduction from* RAW *to* CODA.

(i)

	BUT		MIT		JHU		NDSC		AALTO		Human	
	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$
RAW	56.28	48.91	46.80	38.03	43.11	33.28	42.67	33.69	39.63	30.61	15.08	4.84
RAW$_{norm}$	55.29	48.84	45.57	37.79	41.88	33.33	41.48	33.69	38.09	30.27	13.04	4.18
RAW$_{char}$	33.02	22.29	27.34	18.02	29.41	19.42	24.52	15.31	22.32	14.15	5.94	1.32
CODA	55.16	49.01	45.36	38.08	41.74	33.48	41.20	33.70	37.74	30.30	12.65	4.38
ATB	53.06	46.23	43.73	36.03	40.87	32.56	39.46	31.94	35.80	28.23	11.12	3.80
D3	50.22	43.28	41.19	33.56	38.69	30.49	36.95	29.50	33.56	26.22	10.19	3.39
LEMMA	49.19	43.14	39.62	32.67	36.98	29.14	35.57	28.43	32.69	25.77	11.05	4.00
RAW - RAW$_{norm}$	*0.99*	*0.07*	*1.23*	*0.24*	*1.23*	*-0.05*	*1.19*	*0.00*	*1.54*	*0.34*	*2.04*	*0.66*
RAW - CODA	*1.12*	*-0.10*	*1.44*	*-0.05*	*1.37*	*-0.20*	*1.47*	*-0.01*	*1.89*	*0.31*	*2.43*	*0.46*
% relative reduction	*1.99*	*-0.20*	*3.08*	*-0.13*	*3.18*	*-0.60*	*3.45*	*-0.03*	*4.77*	*1.01*	*16.11*	*9.50*

(ii)

	Metric Based Rankings											
	BUT		**MIT**		**JHU**		**NDSC**		**AALTO**		**Human**	
Average / Absolute Rank	**4.95 / 6**		**4.03 / 5**		**3.93 / 4**		**3.49 / 3**		**2.96 / 2**		**1.14 / 1**	
	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$	WER$_{avg}$	WER$_{mr}$
RAW	6	6	5	5	4	3	3	4	2	2	1	1
RAW$_{norm}$	6	6	5	5	4	3	3	4	2	2	1	1
RAW$_{char}$	6	6	4	4	5	5	3	3	2	2	1	1
CODA	6	6	5	5	4	3	3	4	2	2	1	1
ATB	6	6	5	5	4	4	3	3	2	2	1	1
D3	6	6	5	5	4	4	3	3	2	2	1	1
LEMMA	6	6	5	5	4	4	3	3	2	2	1	1

with the ability to drag and drop the output to indicate the rank. The judges were instructed to read the human references first, then read all of the system outputs. After that they were to rank the system outputs from 1 (best) to up to 6 (worst). Duplicates received the same rank, and no rank skipping was utilized.

We used 100 sentences to evaluate inter-annotator agreement among the judges who were split in two groups with equal number of sentences. Of all the 600 ranking decisions (100 sentences, 6 systems) done twice, the two groups exactly agreed on the rank 46% of the time. The pairwise correlation between the two annotation groups of the average rank per system over the 100 sentences is 97.93%. We take these results to be a strong indicator that this task is a reasonable annotation task and that the human references can be used in our validation.

4. Experimental Results

4.1. Experimental settings

WER metric variant. We compare three WER-based metrics: Single-reference WER (WER$_{one}$), Average WER (WER$_{avg}$), and Multi-reference WER (WER$_{mr}$). Both WER$_{avg}$ and WER$_{mr}$ use multiple references. For WER$_{avg}$, the independent single-reference WER scores are computed, then averaged. For WER$_{mr}$, all references are used to compute the score in a similar manner to the multi-reference BLEU metric used in machine translation evaluation [8, 9].

Text representation. We compare seven text representations:

- RAW and RAW$_{norm}$, refer to raw text, and simple Alif/Ya/Ta-Marbuta normalized text. The normalization removes distinctions within three sets of characters that are often written inconsistently in DA: Alif forms (ا *A*, أ >, إ <, آ |), Ya forms (ي *y*, ى *Y*), and Ta-Marbuta forms (ة *p*, ه *h*).
- RAW$_{char}$ is raw text tokenized at the character level.
- CODA is the Conventional Orthography for Dialectal Arabic.
- ATB and D3 are two commonly used Arabic NLP tokenization schemes that separate clitics from the base word. D3 separates all clitics; while ATB keeps the definite article ال+ *Al+*, e.g., والكتاب *wAlktAb* 'and the book' is و+ ال+ كتاب *w+ Al+ ktAb* in D3, and و+ الكتاب *w+ Alk-tAb* in ATB.

Table 3: *Correlation (%) of a number of metrics against human rankings on the validation set at system level.*

	Average Human Rank vs Metric Score			*Absolute Human Rank vs Metric Rank*		
	WER_{avg}	WER_{mr}	WER_{one}	WER_{avg}	WER_{mr}	WER_{one}
RAW	98.53	98.24	98.46	**100.00**	94.29	96.19
RAW_{norm}	98.59	98.34	98.53	**100.00**	94.29	96.19
RAW_{char}	98.53	98.51	98.50	94.29	94.29	94.29
CODA	98.66	98.41	98.62	**100.00**	94.29	98.10
ATB	98.94	98.83	98.92	**100.00**	**100.00**	**100.00**
D3	**99.06**	**98.93**	**99.04**	**100.00**	**100.00**	**100.00**
LEMMA	99.04	98.71	98.99	**100.00**	**100.00**	**100.00**

- LEMMA is the lemma form that abstract from all inflectional morphology.

Computing correlations. Each experimental setting pairs a WER-based metric with a text representation. The results of the metric-text-representation pairs are evaluated by the correlation of the rankings they assign with human judgment ranking.

There are a number of ways to compute the correlations in our setup varying in terms of system granularity and types of scores and ranks used. The sentence-level human ranks can be used to compute an average human rank of the system, as well as an absolute system human rank (derived from the average human rank per sentence). The system metric scores can similarly be converted to absolute ranks at system level.

We consider two settings that we deem most meaningful: the correlation of average human rank and metric score, and the correlation of absolute human rank with absolute metric rank (Table 3). We present these results for WER_{one}, WER_{avg}, and WER_{mr}. For the WER_{one} case, we present the average of the correlations on the independent single references.

4.2. Metrics and text representation scores

The top half of Table 2 (i) presents the six system scores across WER_{avg} and WER_{mr} for our seven representations. The greatest decrease in WER from the RAW baseline is with RAW_{char}; however, RAW_{char} is the worst in rank prediction. All of the non-RAW representations do better. The drop in WER because of CODA compared to RAW_{norm} is comparable to the human-on-human results in Figure 1. The bottom half of Table 2 (ii) shows the human average rank, absolute rank and the rank suggested by the specific metric-text-representation pairings. The cells in red mark instances of disagreement between human rank and metric rank. The Human "system" is ranked first by all the metrics-text-representation pairs. The best and worst automatic systems are identified correctly by all metric-text-representation pairs. Only ATB, D3 and LEMMA perfectly rank all systems using both WER_{avg} and WER_{mr}.

4.3. Correlations with human judgment

Tables 3 summarizes the correlation scores for the various systems. Our results show that D3 is the best representation to use. It almost bridges the gap between single reference WER and multi-reference WER. The right side of Table 3 reflects the rank pattern in the bottom half of Table 2 (ii).

While these correlations are very high and in a small range, it is important to highlight the following. While the WER metrics using multiple references (WER_{avg} and WER_{mr}) may give lower WER scores per system, their system ranking power is close to the single reference WER metrics. Furthermore, the use of text processing techniques, from spelling conventionalization to lemmatization seem to increase the correlation with human judgment. We note that the single-reference WER setups can only reach perfect rank correlation when using the more abstract representations. This suggests that such representations can be a better solution than using multiple references, which are expensive to create.

Raw Text		CODA		D3		Lemma	
Human	**Machine**	**Human**	**Machine**	**Human**	**Machine**	**Human**	**Machine**
ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*
فيش *fy$*	فيش *fy$*	فيش *fy$*	فيش *fy$*	فيش *fy$*	فيش *fy$*	فِيْش *fy$*	فِيْش *fy$*
حتوافق *HtwAfq*	هتوافق *htwAfq*	حتوافق *HtwAfq*	حتوافق *HtwAfq*	ح+ توافق *H+ twAfq*	ح+ توافق *H+ twAfq*	وافق *wAfq*	وافق *wAfq*
حتوافق *HtwAfq*	أتواجد *>twAjd*	حتوافق *HtwAfq*	أتواجد *>twAjd*	ح+ توافق *H+ twAfq*	أتواجد *>twAjd*	وافق *wAfq*	تواجد *twAjd*
إحتمالات *<HtmAlAt*	احتمالات *AHtmAlAt*	احتمالات *AHtmAlAt*	احتمالات *AHtmAlAt*	احتمالات *AHtmAlAt*	احتمالات *AHtmAlAt*	إحْتِمال *AHtmAl*	إحْتِمال *AHtmAl*
إنك *<nk*	إنك *<nk*	إنك *<nk*	إنك *<nk*	إن +ك *<n +k*	إن +ك *<n +k*	إنْ *<n*	إنْ *<n*
ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*	ما *mA*
توافقش *twAfq$*	توافقش *twAfq$*	توافقش *twAfq$*	توافقش *twAfq$*	توافق +ش *twAfq +$*	توافق +ش *twAfq +$*	وافق *wAfq*	وافق *wAfq*
غالبا *gAlbA*	غالبا *gAlbA*	غالبا *gAlbA*	غالبا *gAlbA*	غالبا *gAlbA*	غالبا *gAlbA*	غالِب *gAlb*	غالِب *gAlb*
مش *m$*	مش *m$*	مش *m$*	مش *m$*	مش *m$*	مش *m$*	مِش *m$*	مِش *m$*
واردة *wArdp*	وارد *wArd*	واردة *wArdp*	وارد *wArd*	و+ أرد +ه *w+ >rd +h*	و+ أرد *w+ >rd*	رَدّ *rd*	رَدّ *rd*
المهم *Almhm*	منه *mnh*	المهم *Almhm*	منه *mnh*	ال+ مهم *Al+ mhm*	من +ه *mn +h*	مُهِمّ *mhm*	مِن *mn*
WER = 41.67%		WER = 25.00%		WER = 26.32%		WER = 16.67%	

Figure 2: *Example of different types of disagreements between a human reference and an ASR output (machine) in Raw text, CODA and lemma representations.*

4.4. Example

Figure 2 presents an example of the type of disagreements that CODA, D3 and LEMMA resolve. Although the lemma is incorrect in cases that increases agreement, the matching is plausible and felicitous. D3 allows for token (subword) matches (see the orange cells).

5. Conclusion and Future Work

We presented a number of evaluation metrics for dialectal ASR, varying in terms of text representations and number of required references. We validated these metrics by comparing their correlations against a human-ranked 1,000 utterance set for six systems. Our results show that the use of morphological abstractions and spelling normalization produces systems with higher correlation with human judgment. These representations can be a better solution than using multiple references, which are expensive to create.

In the future, we plan to make use of these metrics as part of the optimization of hyper parameters in dialectal ASR systems. Another line of research is to explore using the abstract sub-word morphological representation to reduce the out-of-vocabulary similar to [22]. We also plan to expand the human judgment validation set to cover the full MGB-3 set. Furthermore, we plan to conduct a large-scale study on other dialects beside Egyptian Arabic.

6. References

[1] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[2] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU '17, Okinawa, Japan, 2017.

[3] A. Ali, "Multi-dialect Arabic broadcast speech recognition," Ph.D. dissertation, The University of Edinburgh, 2018.

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

[5] R. Rosenfeld and P. Clarkson, "CMU-Cambridge statistical language modeling toolkit v2," 1997.

[6] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU '97, Santa Barbara, California, USA, 1997.

[7] A. Ali, P. Nakov, P. Bell, and S. Renals, "Werd: Using social text spelling variants for evaluating dialectal speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU '17, Okinawa, Japan, 2017.

[8] A. Ali, W. Magdy, P. Bell, and S. Renals, "Multi-reference WER for evaluating ASR for languages with no orthographic rules," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU '15, Scottsdale, Arizona, USA, 2015.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA, 2002.

[10] N. Habash, A. Soudi, and T. Buckwalter, "On Arabic transliteration," in *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, A. van den Bosch and A. Soudi, Eds. Springer, Netherlands, 2007.

[11] N. Habash, M. Diab, and O. Rambow, "Conventional orthography for dialectal Arabic," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey, 2012.

[12] R. Eskander, N. Habash, O. Rambow, and N. Tomeh, "Processing spontaneous orthography," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[13] M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash, and R. Eskander, "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, 2014.

[14] A. Bies, Z. Song, M. Maamouri, S. Grimes, H. Lee, J. Wright, S. Strassel, N. Habash, R. Eskander, and O. Rambow, "Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus," in *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar, 2014.

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

[16] N. Y. Habash, *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers, 2010, vol. 3.

[17] K. Darwish, "Arabizi detection and conversion to Arabic," in *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar, 2014, pp. 217–224.

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

[19] A. Erdmann, N. Zalmout, and N. Habash, "Addressing noise in multidialectal word embeddings," in *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.

[20] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, ser. SLT '2016, San Diego, California, USA, 2016.

[21] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New York City, NY, 2006.

[22] P. Smit, S. R. Gangireddy, S. Enarvi, S. Virpioja, and M. Kurimo, "Character-based units for unlimited vocabulary continuous speech recognition," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, ser. ASRU '17, Okinawa, Japan, 2017.