Revisiting Common Assumptions about Arabic Dialects in NLP

Amr Keleg, Sharon Goldwater, Walid Magdy
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
a.keleg@sms.ed.ac.uk, {sgwater,wmagdy}@inf.ed.ac.uk

Abstract

Arabic has diverse dialects, where one dialect can be substantially different from the others. In the NLP literature, some assumptions about these dialects are widely adopted (e.g., "Arabic dialects can be grouped into distinguishable regional dialects") and are manifested in different computational tasks such as Arabic Dialect Identification (ADI). However, these assumptions are not quantitatively verified. We identify four of these assumptions and examine them by extending and analyzing a multi-label dataset, where the validity of each sentence in 11 different country-level dialects is manually assessed by speakers of these dialects. Our analysis indicates that the four assumptions oversimplify reality, and some of them are not always accurate. This in turn might be hindering further progress in different Arabic NLP tasks.

1 Introduction

Arabic has more than 420 million speakers, and is the official language of more than 22 countries, making it the sixth most spoken language worldwide (Bergman and Diab, 2022). Arabic speakers distinguish between two varieties of the language. Modern Standard Arabic (MSA) is the language of literary work, official documents, and newspapers. MSA has standardized orthography, is taught in schools, and is mostly perceived as a shared variety across Arab countries. Conversely, local dialectal varieties—known as Dialectal Arabic (DA)—are mostly spoken, yet have recently become more written with the rise of social media platforms, despite not having a standardized orthography. These local varieties could differ from MSA and each other in phonology, morphology, syntax, and semantics. Different levels are used to group the varieties of DA as varieties spoken into (a) 5-6 macro-regions, (b) >20 countries, and (c) >100 cities/provinces.

Variation also exists within the same dialect. To quantify this variation, Keleg et al. (2023) introduced the Arabic Level of Dialectness (ALDi) metric, defined as how divergent a sentence is from MSA. ALDi is operationalized as a continuous score between 0 (MSA) and 1 (Highly Dialectal), on the level of sentence-like units.

Successful Arabic NLP systems need to handle all of these types of variation, yet some literature rests on certain assumptions about Arabic dialect variation. In this paper, we identify three common assumptions that were progressively adopted by the Arabic NLP community, in addition to a fourth one that was recently introduced.[1] The assumptions impact different aspects such as distinguishing between the varieties of DA (**Asm. 1**, **Asm. 2**, and **Asm. 4**), and dialectal samples curation (**Asm. 3**). However, their validity is neither backed by enough linguistics studies nor quantitively assessed, making them anecdotal. While they were useful in achieving progress in tasks like Arabic Dialect Identification (ADI)[2], inaccuracies in these assumptions might hinder further progress. Our analysis focuses on the text modality, but the findings could apply to the speech modality. It could also benefit linguists studying the Arabic varieties. We systematically examine the assumptions below:

Asm. 1 A DA sentence is usually valid in only one regional dialect.

Asm. 2 Only short sentences can be valid in multiple dialects.

Asm. 3 Distinctive dialectal words (e.g., برشة /bršħ/ for Tunisian Arabic) can be curated to infer the dialect of sentences containing any of them.

Asm. 4 For a sentence valid in multiple dialects, speakers of these dialects consistently provide similar ratings of the sentence's level of dialectness.

[1] Limitations of the 4 assumptions are discussed qualitatively in the literature but are ignored or perceived as minor.

[2] As of the 15th of December 2024, 618 papers on Semantic Scholar (Jones, 2015) match "Dialect Identification", out of which 173 (≈28%) match "Arabic Dialect Identification". However, ADI is still unsolved (Abdul-Mageed et al., 2024).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3309–3327

In our analysis[3], we used 978 DA sentences geolocated to 14 different Arab countries. 33 annotators from 11 Arab countries (3 each) labeled each sentence for **(a)** validity in the annotator's country-level dialects and **(b)** Arabic Level of Dialectness (ALDi). We find that >56% of the dataset is valid in multiple regional dialects, showcasing that ADI is a multi-label classification task (i.e., each sentence should be assigned multiple labels, not a single one). The sentence's ALDi correlates better with the number of its valid dialects than its length. Moreover, lists of dialectal words are not always distinctive of their presumed dialects. Lastly, the ALDi ratings assigned by speakers of different regional dialects can significantly vary, for sentences valid in these dialects.

2 Background

In this section, we describe how the four assumptions were progressively adopted.

2.1 The groupings of Arabic Dialects

Along the vast geographical area over which Arabic speakers are distributed, different varieties of DA are spoken. Varieties spoken within geographically proximate areas are commonly grouped into regional dialects. An example of such groupings is: the Levant (Lebanon, Jordan, Palestine, Syria), Nile Basin (Egypt, Sudan), Gulf (Saudi Arabia, Oman, Qatar, Bahrain, United Arab Emirates, Iraq), Gulf of Aden (Yemen, Djibouti, Somalia), and Maghreb (Morocco, Tunisia, Algeria, Mauritania, Libya).[4] Regional groupings recognize the within-region similarities while assuming minimal overlap between the regional varieties.

Regional-level Dialects Early efforts in ADI used single-label classification to distinguish between a subset of the regional varieties, including MSA as an independent variety/class (Biadsy et al., 2009; Zaidan and Callison-Burch, 2011). This adoption of single-label classification implicitly accepts **Asm. 1** at the regional level; i.e., that sentences are usually only valid in one regional dialect. Three follow-up papers did back off from this assumption by introducing a new class (*General*) for sentences that are valid in multiple regional dialects (Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Zaidan and Callison-Burch, 2014). The last of these papers found that *General* class represented $\approx 6.3\%$ of the total annotations in their dataset, demonstrating how the regional dialects are not fully distinguishable from each other. However, the authors also noted that some annotators wrongly selected the *General* class when they could not decide the dialect of the sentence, while others labeled some sentences as only valid in their native dialects although these sentences are valid in other dialects.

Despite these hints of additional complexity, overlap between the regional dialects was ignored in annotating further datasets (Bouamor et al., 2014; Salama et al., 2014; Huang, 2015; Malmasi et al., 2016; Zampieri et al., 2017, 2018; El-Haj et al., 2018; Alsarsour et al., 2018; Abu Farha et al., 2021).[5] A few papers acknowledge this limitation of their datasets, providing examples of sentences that are valid in multiple regional dialects (Malmasi et al., 2016; Lulu and Elnagar, 2018; Salloum, 2018; El-Haj, 2020), or valid in both MSA and a regional dialect[6] (El-Haj et al., 2018), but the continued use of single-label annotation implies that these cases are thought to be a small minority.

Country-level Dialects Grouping dialects into regions abstracts differences between the dialects spoken within each region (Shon et al., 2020; Althobaiti, 2020; Messaoudi et al., 2022), such as those between Egyptian and Sudanese Arabic (Abdul-Mageed et al., 2018), or between the dialects of the Levant (Abu Kwaik et al., 2018). Therefore, more fine-grained sets of labels were proposed for the task of ADI. Country-level ADI is the most common setup (Abu Kwaik et al., 2018; Shon et al., 2020; Abdul-Mageed et al., 2022, 2023), with some datasets targeting both country-level and province/city-level ADI (Abdul-Mageed et al., 2018; Salameh et al., 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020a, 2021).

Country-level ADI has still been modeled as a single-label classification task. This is problematic as any overlap existing on the regional level will still exist when these regions are divided into countries. Moreover, similar country-level dialects of the same region are expected to overlap. Hence, it has been found that many errors of the country-level ADI models are caused by confusing dialects

[3] We release our code at: https://github.com/AMR-KELEG/MLADI-assumptions-revisiting

[4] A canonical grouping of the Arabic dialects does not exist (Habash, 2010; Abdul-Mageed et al., 2018).

[5] See §A for a discussion on regional ADI performance.

[6] Some phonological differences are lost in text, making some sentences plausible in both MSA and a variety of DA.

spoken in neighboring countries, most of which would belong to the same region (Biadsy et al., 2009; Salameh et al., 2018; Talafha et al., 2019; Samih et al., 2019; Ragab et al., 2019; Přibáň and Taylor, 2019; Ghoul and Lejeune, 2019; Eltanbouly et al., 2019; Abu Kwaik and Saad, 2019; Dhaou and Lejeune, 2020; Talafha et al., 2020; Aloraini et al., 2020; Abdelali et al., 2021; AlKhamissi et al., 2021; El Mekki et al., 2021; Jamal et al., 2022; Khered et al., 2022; Attieh and Hassan, 2022).

Sentence Length and ADI Most ADI datasets use sentence-like units (e.g., tweets). A common belief (**Asm. 2**) is that most multi-label samples are very short. Since most NLP models would struggle with these short sentences, holding this belief might explain why ADI has continued to be modeled as a single-label classification task.

2.2 Dialectal Lexical Cues

Although dialects differ at many linguistic levels (phonological, lexical, syntactic), one of the easiest types of cues to identify in text is lexical cues (Kaye and Rosenhouse, 1997). These cues are *distinctive* of a particular dialect if they are not shared with other dialects. Some papers provide qualitative examples of these cues like هطعش (/hTʕš/ - eleven)[7] for Yemeni (Al-Shargi et al., 2016) and برشة (/baršħ/ - a lot) for Tunisian (McNeil, 2018; Abdelali et al., 2021).

Distinctive cues have been widely used to build DA datasets. To this end, ad-hoc lists of lexical cues were compiled to collect dialectal samples from websites or social media platforms. These lists were either directly used (Al-Sabbagh and Girju, 2012; Alshutayri, 2017; Alshargi et al., 2019), or first validated by speakers of different dialects to ensure their distinctiveness (Almeman and Lee, 2013; Zaghouani and Charfi, 2018; Alsarsour et al., 2018; Mubarak, 2018).

It is acknowledged that the diversity of the curated samples is limited by the lists of cues (Abdul-Mageed et al., 2020b). However, the precision and distinctiveness of these cues are assumed to be high without quantitatively measuring them (**Asm. 3**), which we revisit in this paper.

2.3 Differences in ALDi Perceptions

The concept of having different levels of dialectness was noted decades ago (Badawi, 1973; Parkinson, 1991). In NLP, two papers designed guidelines for rating the level of dialectness of sentences (Habash et al., 2008; Zaidan and Callison-Burch, 2011). After that, however, the concept was ignored until Keleg et al. (2023) proposed fine-tuning a BERT-based model to automatically quantify it as a score in [0, 1] on sentence-like units. They found that some sentences can be considered as being closer to MSA or DA based on how an annotator attempts to pronounce them. Therefore, they embrace the variation in the human annotations by averaging them to obtain gold standard ALDi scores. However, this overlooks the impact of the annotator's native dialect on the provided ALDi ratings (**Asm. 4**).

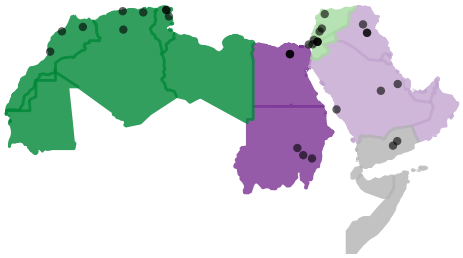


Figure 1: A map of the Arab world. The black dots indicate the provinces/cities from which the annotators originate. Regional dialects (Maghreb, Nile Basin, Levant, Gulf, Gulf of Aden) are encoded as different colors according to the groupings of Baimukan et al. (2022).

3 Data

For our analysis, we release an extended version of the NADI 2024 dataset (Abdul-Mageed et al., 2024), that we call the ***MLADI*** (Multi-label ADI) dataset.[8] The original NADI 2024 dataset has 1,120 tweets, of which only 70 were automatically identified as MSA and 1,050 as DA. The DA samples' geolocations are uniformly distributed across the 14 most populated Arab countries, excluding Somalia, for which data is not sufficiently abundant. 27 annotators were recruited from 9 Arab countries (3 each): Algeria, Morocco, Tunisia, Egypt, Sudan, Palestine, Syria, Iraq, and Yemen. For each sample in the dataset, the annotators **(a)** identified if a speaker of one of their country-level dialects could have authored the tweet. If an annotator answered (a) as yes, then the sentence is also **(b)** rated for its ALDi as MSA (L0), Colloquial-influenced MSA (L1), Normal Colloquial (L2), or Informal (or Vulgar) Colloquial (L3).

[7] Transliteration follows HSB scheme (Habash et al., 2007).

[8] We release an accompanying ADI leaderboard at: https://huggingface.co/spaces/AMR-KELEG/MLADI

The dataset creators provided us with the annotated samples and the individual annotator labels, which we used to study the aforementioned assumptions. In addition, we recruited 3 annotators from each of Jordan and Saudi Arabia to extend the dataset's labels, using the same annotation guidelines as in Abdul-Mageed et al. (2024). This improves the dataset's coverage of the different Arab dialects, especially Gulf Arabic. Figure 1 shows the annotators' cities/provinces of origin.

The Interannotator Agreement scores (see §B) for the two new dialects are similar to the ones reported for the NADI 2024 dataset. Following the NADI 2024 paper, we use majority voting to identify the validity of each tweet in each of the 11 country-level dialects, and for ALDi, we transform the ratings from discrete levels (L0, L1, L2, L3) into numeric values $(0, \frac{1}{3}, \frac{2}{3}, 1)$. A sentence's ratings, for the dialects in which the sentence is valid (according to the majority voting), are averaged to estimate a dialect-agnostic ALDi score.

4 Analysis

In this section, we investigate each of the four assumptions listed in §1, using 978 out of the 1,050 DA samples, after discarding 72 samples that are not labeled as valid in any of the 11 considered country-level dialects.

4.1 Asm. 1 - Arabic Dialects Rarely Overlap

At least 28 different ADI datasets assign a single regional/country-level dialect to each sentence (Keleg and Magdy, 2023). Single-label classification was shown not to be suitable for country-level ADI both qualitatively (Kchaou et al., 2019; Touileb, 2020; Bayrak and Issifu, 2022; Khered et al., 2022) and quantitatively (Keleg and Magdy, 2023; Olsen et al., 2023; Abdul-Mageed et al., 2024). However, single-label classification might still be thought of as suitable for ADI on the level of regional dialects, under the assumption that they rarely overlap.

Method Using the regional grouping proposed by Baimukan et al. (2022), we form 5 regional-level validity labels from the 11 country-level labels as follows: **1) Nile Basin (NL):** Egypt, Sudan, **2) Gulf (GL):** Iraq, Saudi Arabia, **3) Gulf of Aden (AD):** Yemen, **4) Maghreb (MG):** Tunisia, Algeria, Morocco, and **5) Levant (LV):** Jordan, Palestine, Syria. A sentence is valid in a regional dialect if it is valid in at least one of the considered region's countries. Afterward, we count the number of regional dialects in which each sentence is valid.

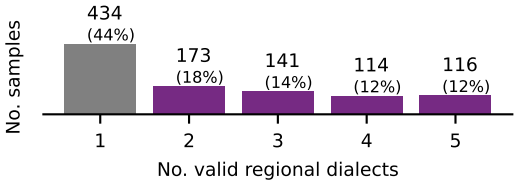


Figure 2: The histogram of the number of valid dialects on the regional level. Only 44% of the DA samples are confined to single-region dialects.

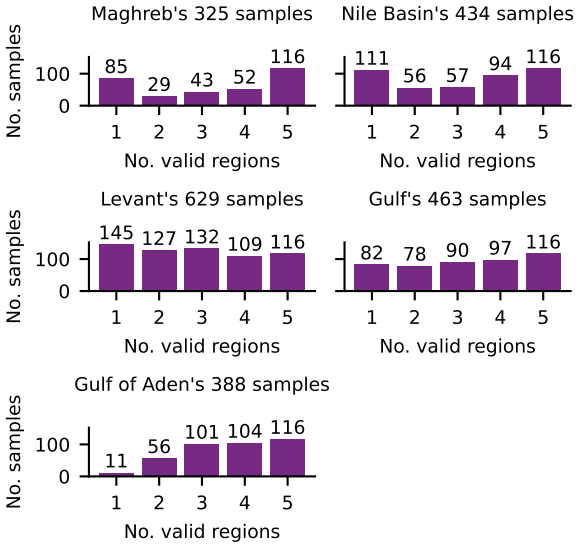


Figure 3: The total number of valid regional dialects for each region's valid samples. **Note:** The regions' samples are not mutually exclusive (e.g., the same 116 samples valid in the 5 regions are in all distributions).

Results A majority 56% of sentences (544 in total) are valid in multiple regional dialects, as shown in Figure 2. This large cross-regional overlap exists despite the fact the MSA samples were discarded. Notably, 116 of these DA samples (a non-negligible $\sim$ 12%) are valid in all regional dialects.

Further Analysis Unlike the other dialects, the Gulf of Aden (represented by Yemen) has only 11 single-region samples as per Figure 3. Hence, it might not be prominently different from some of the subdialects spoken in other regions, challenging the recognition of *Gulf of Aden* as a regional dialect (Habash, 2010; Abdul-Mageed et al., 2018).

More broadly, Figure 3 shows that the Levant, Gulf, and Gulf of Aden have a substantial number of samples shared with other regional dialects, with Levantine sharing more than the other two dialects. Looking at the distribution of the multi-region samples in Figure 4, a large number of the 2-region samples are between pairs of these three regions

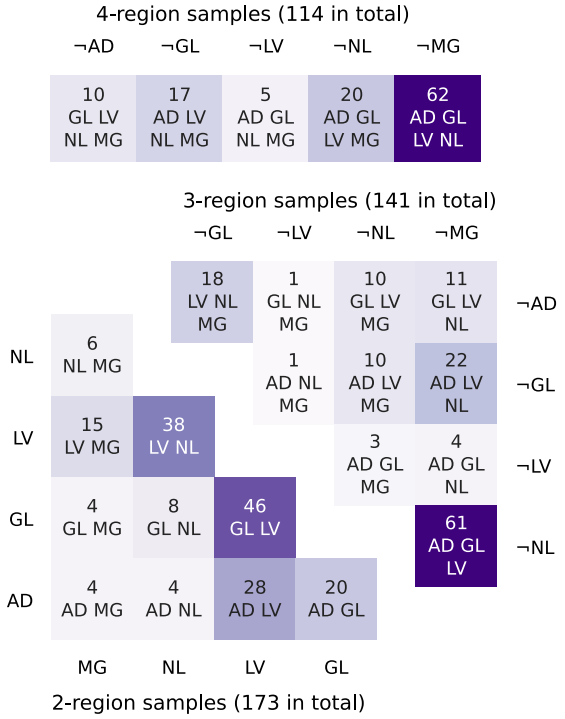


Figure 4: The distribution of the 2-region, 3-region, and 4-region samples across the different combinations. Each combination has its regions indicated in its respective cell. **Note:** GL/¬GL means valid/not valid in Gulf.

(e.g., 46 valid in GL and LV, 20 valid in AD and GL) and a majority of 61 samples of the 3-region ones are valid in these regions. Additionally, LV has a substantial number of 38 samples shared with NL, 15 shared with MG, and 18 shared with both. This explains how LV shares more samples with other dialects than GF and AD.

For the remaining two dialects, both share fewer samples with other dialects, with NL sharing more samples than MG. 62 samples (a majority of the 4-regions samples) are valid in all regions but MG. This is a sign of the dichotomy between the Eastern dialects of Arabic spoken in the Maghreb and the other dialects spoken in the West of the Arab world (Kaye and Rosenhouse, 1997). Still, MG shares more with other dialects than previously assumed.

Implications Substantial overlap exists between the regional dialects, which contradicts the general perception that they are distinguishable from each other. As previously mentioned, this overlap will still exist when the regions are split into countries as shown in §C. Hence, ADI is a multi-label task on both the regional and country levels.

Classifying *Gulf of Aden* as a distinct regional variety requires reevaluation, given the limited number of samples only valid in this region. Similarly, dialectal categorizations that are not based on the country borders could be considered.[9]

4.2 Asm. 2 - Only Short Sentences' Dialects are Ambiguous

In the context of ADI, sentence length is discussed from two points of view (POVs). *POV #1* explicitly mentions that the dialect of extremely short speech segments/text sentences can be ambiguous. Hence, it is infeasible for humans, and consequently machines, to assign a single dialect to these segments (Alorifi, 2008) and sentences (El-Haj et al., 2018; Alsarsour et al., 2018; Abu Kwaik and Saad, 2019; Althobaiti, 2022). *POV #2* empirically finds that the longer the segment/sentence gets, the higher the performance of a single-label ADI system is, for speech (Biadsy et al., 2009; Shon et al., 2020) and text (Zaidan and Callison-Burch, 2014; Salameh et al., 2018; AlKhamissi et al., 2021; Abdelali et al., 2021; Bayrak and Issifu, 2022). This can be attributed to a decline in dialect ambiguity as sentences get longer.

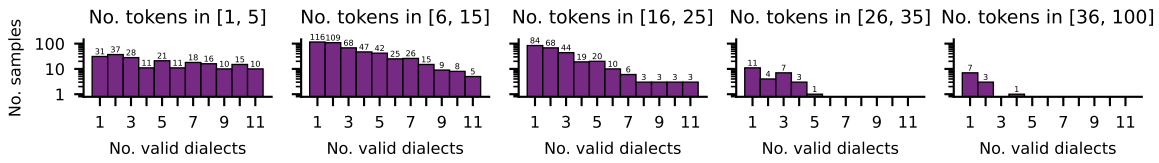
Method We examine the assumption by computing Spearman's correlation between the sentence length (as the number of tokens) and the number of valid dialects on the country level. Additionally, we study the histograms of the number of valid dialects for five different ranges of sentence lengths.

Results According to Figure 5a, the majority of trivially short sentences are valid in multiple dialects as per *POV #1*. However, *POV #1* overlooks the large number of moderately long sentences (16-25 tokens) that are also valid in multiple dialects. Additionally, despite long sentences being valid in a smaller number of dialects, confirming *POV #2*, there is only a weak negative Spearman's correlation coefficient (-0.28) between the sentence length and its number of valid dialects.

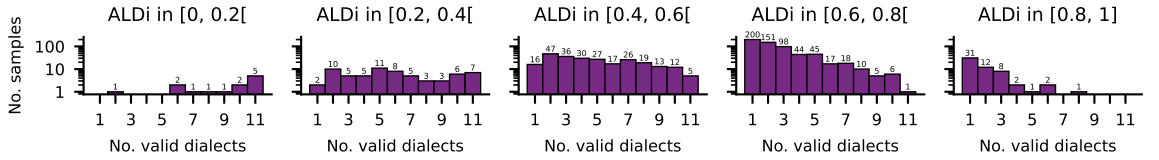
Further Analysis On replicating the analysis by replacing the sentence length with the ALDi score, a stronger negative correlation (-0.52) is realized.[10] Figure 5b also indicates that sentences of ALDi scores < 0.2 are generally valid in most of the dialects. Samples with ALDi scores $\in [0.2, 0.4[$ seem to be evenly probable across the different number of validity labels. The distribution then

[9] Glottolog and Ethnologue recognize 37 and 28 Arabic dialects, respectively.

[10] A coefficient of -0.45 is realized when replacing the aggregated manually-assigned ALDi scores with ones automatically estimated using the Sentence-ALDi model (Keleg et al., 2023).



(a) Sentence length (measured as the number of tokens). **Note:** $\rho(Sentence\ Length, No.\ valid\ dialects) = -0.28$



(b) ALDi scores (averaged across all ratings). **Note:** $\rho(ALDi, No.\ valid\ dialects) = -0.52$

Figure 5: The distribution of the sentences (log scale) and the number of valid country-level dialects according to different ranges of sentence length **(a)** and ALDi scores **(b)**. **Note:** Since the MSA samples were automatically discarded from our analysis dataset, there are very few samples with low ALDi scores ($\in [0, 0.2]$). However, the histogram of this bin is expected to be left-skewed (i.e., MSA samples are expected to be valid in all dialects).

shifts to be more and more right-skewed for the subsequent ranges of ALDi scores.

Implications Previous assumptions about sentence length are either incomplete (*POV #1*) or not sufficiently accurate (*POV #2*). Moreover, a sentence's ALDi score correlates moderately with the number of dialects in which it is valid, making it a better predictor than sentence length. As a proxy of a sentence's number of valid dialects, ALDi could guide the predictions of a multi-label ADI system.

4.3 Asm. 3 - Dialects' Distinctive Lexical Cues

Method For each of DART's (Alsarsour et al., 2018) and DIAL2MSA's (Mubarak, 2018) lists of regional-level distinctive cues, we identify sentences of our dataset that match at least one of the lexical cues.[11] We normalize the sentences and lists of cues to handle common typos/ dialectal variations of the same characters (e.g., ة is normalized to ه and إ, أ, آ are normalized to ا) (Kholy and Habash, 2012; Darwish and Magdy, 2014). Exact matching is then used between the lexical cues and the whitespace tokenized sentences' tokens.

For each dialect, we report the number of samples matching at least one of its distinctive cues (M). Then, we count the number of matching samples manually annotated as valid in this dialect (M_{Val}), and the number of matching samples that are only (i.e., exclusively) valid in this dialect (M_{Exc}). Precision (P), Distinctiveness (D), and Recall (R) of each list are computed as $P = \frac{M_{\text{Val}}}{M}$, $D = \frac{M_{\text{Exc}}}{M}$, and $R = \frac{M_{\text{Val}}}{N_{\text{Val}}}$; where ($N_{Val}$) is the total number of samples valid in the considered dialect.

Adhering to the regional groupings used in both lists, we aggregate the 11 country-level validity labels into the following regions: **1) Egypt**, **2) Iraq**, **3) Gulf:** Saudi Arabia, **4) Maghreb:** Algeria, Morocco, Tunisia, **5) Levant:** Jordan, Palestine, Syria. The dialects of Sudan and Yemen were ignored in both lists, so we considered them as **6) Others**.

Results Table 1 shows the results. The extremely low range of recall values for both manually validated lists confirms that relying on these lists of cues limits the number of matching samples. Conversely, the range of the precision scores is generally high (yet not perfect), except for the cues of Gulf Arabic. The Egyptian Arabic cues have a low precision score (0.6) for DART and extremely low distinctiveness values (0.35 and 0.38) for both lists.

The samples' validity in the Maghreb, Levant, and Gulf regions is only defined by the subset of the region's countries from which we could recruit annotators. Hence, the precision scores for these regions might improve after collecting annotations for more country-level dialects. However, the non-perfect Distinctiveness scores indicate that some cues of these regions are used in other regional dialects, even when the cues were manually validated for their distinctiveness by the lists' creators.

Qualitative Analysis On manually inspecting the matching samples, we found that DART's three matching cues of Gulf Arabic (شنو /šnw/, علامك /ʕlAmk/, مواعين /mwAʕyn/) are indeed dialec-

[11] We could not get access to the lists of (Almeman and Lee, 2013; Zaghouani and Charfi, 2018; Alshargi et al., 2019).

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	60	36	21	287	.60	.35	.13	271	28
IRQ	7	6	6	204	.86	.86	.03	120	7
MGH	21	16	14	325	.76	.67	.05	273	13
LEV	32	29	25	629	.91	.78	.05	240	11
GLF	9	0	0	407	.00	.00	.00	200	3

(a) DART's 5 regional lists.

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	53	43	20	287	.81	.38	.15	28	19
MGH	45	36	31	325	.80	.69	.11	60	26
LEV	38	34	34	629	.89	.89	.05	31	11
GLF	0	-	-	407	-	-	.00	9	0

(b) DIAL2MSA's 4 regional lists.

Table 1: The Precision (*P*), Distinctiveness (*D*), and Recall (*R*) of each region's cues. **Note:** For each region's list, we report the number of samples of our dataset matching any of the cues (*M*) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}), in addition to the total number of valid samples (N_{Val}). The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$).

tal terms that are valid in other regional dialects, hence are not indicative of Gulf Arabic. Additionally, other terms are false friends, having different meanings in MSA and DA varieties, and are not distinctive of a specific dialect in the absence of context. For instance, the terms (ماشي /mAšy/ and حد /Hd/) have the meanings *okay* and *someone* in Egyptian Arabic. However, they have different meanings in MSA (*walking* and *limit*). The MSA sense of these terms could be used in the context of other dialects, as demonstrated in the examples below, which both use the term حد /Hd/ (underlined in the examples). Example (1) uses this term with its Egyptian meaning (*someone*) and is labeled as valid in Egyptian, whereas (2) uses the term with its MSA meaning (*limit*) and is labeled as valid in Algerian and Tunisian. Therefore, the term حد /Hd/ cannot be considered a valid cue to Egyptian Arabic, as assumed in DART.

(1) دا الفرق بين حد اهله عرفوا يربوه وحد تاني منفعش فيه التربيه.
'This is the difference between a well-mannered and a bad-mannered person.'

(2) الي حد الان مازال حظوظ تونس كبيره هاك تراو الفرق الكبار.
'So far, Tunisia still has great chances, this is how big teams are.'

Implications More rigor is needed in building lists of distinctive dialectal words, especially when the curated sentences need to be surely valid in a specific dialect and/or exclusively valid in this dialect. Using a second validation step (e.g., information about the geolocation of the sentence's author) could increase the precision of the dialects assigned based on the cues' associated dialects. However, this does not ensure distinctiveness and further decreases the recall (see §D).

4.4 Asm. 4 - ALDi Perceptions across Dialects

Inspired by earlier work (Zaidan and Callison-Burch, 2011), Keleg et al. (2023) introduced the idea of ALDi prediction as an important task. Two recent datasets provide pairs of sentences with their corresponding aggregated ALDi scores: *AOC-ALDi* (Keleg et al., 2023) and *NADI 2024* (Abdul-Mageed et al., 2024). For the former, three annotations per sentence were sought by randomly assigning the sentences to speakers of different dialects (Zaidan and Callison-Burch, 2011). For the latter, 27 annotators rated the ALDi of each sentence only when it was valid in their country-level dialect. Both datasets used the mean of a sentence's ALDi ratings as its gold-standard ALDi score. The implicit assumption is that ALDi scores do not depend on the annotator's native dialect; however, this has not been empirically validated. We have shown (§4.2) that even sentences with moderate ALDi scores can be valid in multiple dialects, but it is possible that the scores assigned by annotators from those dialects could systematically differ.

Method We compute the Mean Difference (MD) of country-level ALDi scores for each pair of countries. MD is computed for a pair of countries r and c, with N_{rc} sentences valid in both, as

$$\text{MD}(r, c) = \frac{1}{N_{rc}} \sum_{i=1}^{N_{rc}} (\text{ALDi}_r[i] - \text{ALDi}_c[i]),$$

where $\text{ALDi}_r[i]$ and $\text{ALDi}_c[i]$ are the averages of sentence i's ALDi ratings provided by the annotators of r and c respectively.

Results Figure 6 summarizes the results. The top three (orangish) rows indicate that when sentences are valid in one of the Maghreb's countries and another non-Maghrebi country, the annotators from the Maghrebi country rate these sentences to

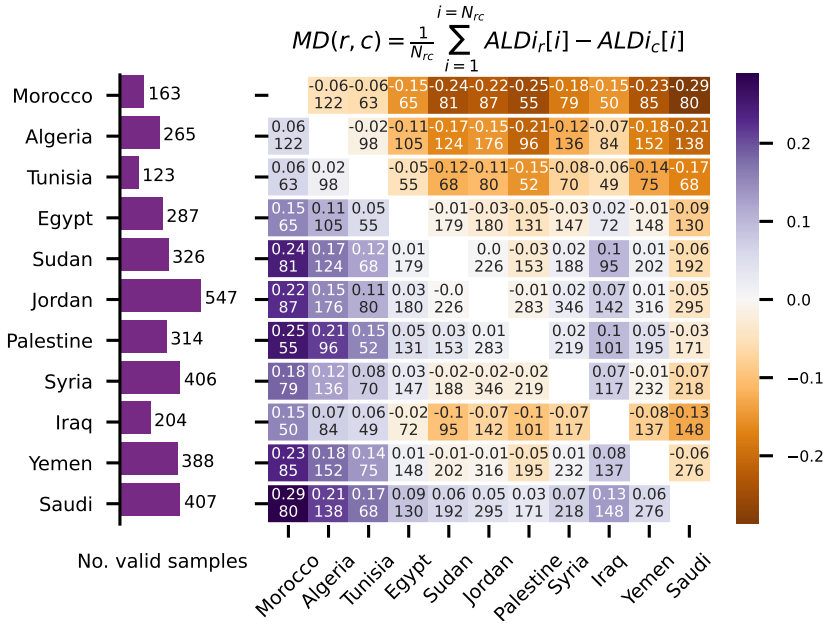


Figure 6: **(Left)** The number of valid samples per country (with countries ordered such that same-region ones are consecutive). **(Right)** Mean difference (MD) of row country's (r) and column country's (c) ALDi scores, for the N_{rc} sentences valid in both (N_{rc} is shown as the bottom number in each cell).

be less dialectal than the non-Maghrebi ones. The difference (e.g., MD(Morocco, Saudi) = -0.29) can be close to $\frac{1}{3}$, which is the difference between two consecutive levels of ALDi ratings $(0, \frac{1}{3}, \frac{2}{3}, 1)$. A similar pattern holds true for Iraq to a lesser extent. Conversely, Saudi annotators assign higher ALDi scores to sentences common with other dialects. Many of the country-level differences are statistically significant, with Standard Errors < 0.035. However, these differences could arise simply because the annotators differ randomly in their mean scores, independent of dialect. So we might see an apparent difference between country groups if we happened to get annotators with higher means in some countries than in other countries. Due to having only three annotators per country, it is not possible to conclusively test for an effect of dialect (separate from annotator) at the country level, although the consistent trends in the visualization are suggestive. Instead, we test for regional-level differences between annotators, as described below. If additional annotations from each country are obtained in the future, a similar test could be used at the country level.

Statistical Analysis We use a one-sided permutation test to assess whether the differences between two groups of annotators (G_A, G_B), of sizes $|G_A|$ and $|G_B|$ respectively, can be attributed to the groups' dialects. First, we compute the MD score between the observed groups' mean ALDi scores (MD_{obs}), for the N_{AB} sentences valid in both groups. A large number of pairs of groups $\{(A', B')\}$ with sizes $|G_A|$, $|G_B|$ are sampled (50k in our case). The pairs of groups (A', B') are formed by random shuffling and distributing all the annotators across two groups. MD scores for each pair are computed for the same N_{AB} sentences.[12] The p-value is the percentage of the shufflings with MDs $\leq$ the observed grouping's mean difference (MD_{obs}).

We consider the annotators of each region as a group, merging Gulf and Gulf of Aden into one region based on the findings of §4.1. Accordingly, we find significant MDs of -0.09, -0.13, -0.14 between the ALDi scores averaged across the annotators of Maghreb against those of Nile Basin, Levant, and Gulf/Gulf of Aden, with p-values of 0.007, 0.00002, and 0.0002, respectively. Similarly, Nile Basin's annotators provide significantly lower ALDi scores than Levantine annotators, with MD of -0.05 (p-value = 0.04). Differences between other pairs are not statistically significant.

Discussion There is a general impression that the Arabic dialects are not equally distant from MSA, with some researchers claiming certain dialects—e.g., Gulf Arabic (Zaidan and Callison-Burch,

[12] In some permutations, we discard the small proportion of sentences that have no ALDi ratings for one of the groups.

2014) and Palestinian Arabic (Kwaik et al., 2018)—are closer to MSA than others, which could explain the MDs we found for samples shared between different countries/regions.

Implications Further analysis is required before taking these MDs as an objective measure of a variety's divergence from MSA. Figure 3 indicates that all regions—except *Gulf of Aden*—have many samples not shared with other regions. Single-region samples could still be highly divergent from MSA. Moreover, people's perception of dialectness is influenced by how they use MSA terms colloquially. For example, both خمر /xmr/ and خمرة /xmrħ/ are valid MSA terms for *wine*. The Holy Qur'an mentions the former, while the latter is more colloquially used in Egypt. Hence, Egyptians might link the first to CA/MSA, and the latter to DA. Consider some MSA lexical items that are shared with dialect D_A but not with dialect D_B. Sentences with these items could be rated as more dialectal by speakers of D_A than D_B. Lastly, sentences valid in multiple varieties could share the same surface form but have different semantics in each variety.

5 Further Implications in NLP

Recent improvements to how the varieties of Arabic are computationally modeled (Keleg et al., 2023; Keleg and Magdy, 2023; Abdul-Mageed et al., 2024) are being used in multiple applications, such as better routing of samples to annotators (Keleg et al., 2024), evaluating the LLMs' dialectal capabilities (Robinson et al., 2025), and building better recommendation systems (Alshabanah and Annavaram, 2025). Hence, validating widely-held assumptions about Arabic could lead to further progress in automatic ADI and many other tasks/applications.

For example, Arabic NLP researchers used manually curated lists of words/phrases to curate data for various applications like compiling dialect-specific pretraining data (Gaanoun et al., 2024), creating datasets for sentiment analysis (Refaee and Rieser, 2014), and offensive text classification (Chowdhury et al., 2020). Therefore, our finding—that some terms share the same orthographic form but have different semantic meanings/senses in various varieties of Arabic—has implications for building datasets for tasks beyond ADI.

Moreover, parallels of the first three assumptions exist beyond Arabic. For example, the overlap between different dialects of the same language has already been noted for other languages such as English, French, and Spanish (Bernier-colborne et al., 2023; Zampieri et al., 2024; Lopetegui et al., 2025). Our findings argue for modeling dialect identification as a multi-label classification task, even on macro-regional levels. In addition, sentence length has been discussed as an important predictor of language identification models' performance (Baldwin and Lui, 2010), especially for closely-related languages and dialects (Tiedemann and Ljubešić, 2012; Blodgett and O'Connor, 2017; Kanjirangat et al., 2022). We show that the conscious *Dialect Level* choice that Arabic speakers make—operationalized as ALDi—is a better predictor of the number of dialects in which a sentence is valid than its length. Speakers of other languages make similar conscious decisions about how much they adhere or diverge from the standard variety of their language (e.g., Shoemark et al., 2017). For these languages, modeling the sentences' divergence from the language's standard variety, as ordinal/quantitative variables, could also provide better predictors of a sentence's validity in multiple dialects than the sentence's length.

6 Conclusion and Moving Forward

We identified four common assumptions regarding Arabic dialects, and systematically studied them by extending the annotations of a previous dataset to cover more country-level dialects. Our analysis shows that these assumptions oversimplify some details that, in turn, impact how tasks are framed, datasets are created, and models are trained.

In particular, our main findings and recommendations are as follows. (1) Arabic dialects overlap considerably at both the country and regional levels, so ADI should be modeled as a multi-label task at both levels. (2) Existing lists of supposedly distinctive lexical cues are less distinctive than previously thought. More rigorous validation is needed for such lists in the future. (3) ALDi scores (but not sentence length) provide a good proxy of a sentence's validity in multiple dialects, which could be used to inform annotation and modeling decisions. Nevertheless, researchers should be aware that speakers of different dialects may systematically differ in their ALDi annotations of the same sentences. (4) Future work should study if sentences with diverging ratings by speakers of different dialects have different semantic meanings in these dialects.

Limitations

This paper revisits some widely-held and mostly unquantified assumptions about the Arabic dialects by extending the annotations of the NADI 2024 dataset to have better coverage of the dialects. Replicating the analysis on other datasets would provide more evidence for the generalizability of our results. Moreover, extending our analysis to cover more country-level dialects might uncover more results than the ones we had when considering 11 country-level dialects. The same applies to using a more granular grouping of the Arabic dialects like different dialects spoken within the same country (e.g., city-level/province-level dialects).

Despite having three annotators per country, our crowdsourced annotators are skewed toward younger age groups and have/are pursuing higher education degrees. Therefore, we acknowledge that our results could be representative of the perceptions of specific demographics within each country.

The analyzed tweets' geolocations are uniformly balanced across 14 different Arab countries, covering a wide range of Arabic dialects. However, we acknowledge that some sub-dialects are not well represented online, as shown by Mohamed Eida et al. (2024) for the Sa'idi Arabic variety of Egypt. Moreover, the data does not have Arabic sentences written in Latin script (known as Arabizi). Arabizi is prominently used in the Maghreb region (Younes et al., 2015), and to a lesser extent in other countries such as Lebanon and Egypt (Tobaili, 2016).

Acknowledgments

We thank Adam Lopez, Nina Gregorio, Burin Naowarat, Yen Meng, Oli Liu, and Sung-Lin Yeh for their comments on an earlier draft of this paper. This work was supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics.

Ethical Considerations

The NADI 2024 dataset, which we extended and used in our analysis, has a few samples with offensive language. Our annotators were asked to provide consent confirming their agreement to annotate these samples at the start of the annotation process. The annotation process we followed was approved by the Research Ethics Committee of the University of Edinburgh, School of Informatics, with reference number 839548.

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

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*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

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

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, 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.

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

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kathrein Abu Kwaik and Motaz Saad. 2019. ArbDialectID at MADAR shared task 1: Language modelling and ensemble learning for fine grained Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 254–258, Florence, Italy. 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).

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another dialectal Arabic corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1300–1306, Portorož, Slovenia. European Language Resources Association (ELRA).

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting MARBERT for improved Arabic dialect identification: Submission to the NADI 2021 shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 260–264, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Khalid Almeman and Mark G. Lee. 2013. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Abdulrahman Aloraini, Massimo Poesio, and Ayman Alhelbawy. 2020. The QMUL/HRBDT contribution to the NADI Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 295–301, Barcelona, Spain (Online). Association for Computational Linguistics.

Fawzi Alorifi. 2008. *Automatic Identification of Arabic Dialects Using Hidden Markov Models*. PhD thesis, University of Pittsburgh.

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

Abdulla Alshabanah and Murali Annavaram. 2025. On using Arabic language dialects in recommendation systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2178–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147, Florence, Italy. Association for Computational Linguistics.

AOO Alshutayri. 2017. Exploring Twitter as a source of an Arabic dialect corpus. *International Journal of Computational Linguistics (IJCL)*, 8(2):37–44.

Maha J. Althobaiti. 2020. Automatic Arabic dialect identification systems for written texts: A survey. *Preprint*, arXiv:2009.12622.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Joseph Attieh and Fadi Hassan. 2022. Arabic dialect identification and sentiment classification using transformer-based models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 485–490, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

As-Said Muhámmad Badawi. 1973. مستويات العربية المعاصرة في مصر. *Levels of Contemporary Arabic in Egypt*. Dar Al-Maarif.

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.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Giyaseddin Bayrak and Abdul Majeed Issifu. 2022. Domain-adapted BERT-based models for nuanced

Arabic dialect identification and tweet sentiment analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 425–430, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 53–61, Athens, Greece. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *Preprint*, arXiv:1707.00061.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. 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*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kareem Darwish and Walid Magdy. 2014. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Ghoul Dhaou and Gaël Lejeune. 2020. Comparison between voting classifier and deep learning methods for Arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 243–249, Barcelona, Spain (Online). Association for Computational Linguistics.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. BERT-based multi-task model for country and province level MSA and dialectal Arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sohaila Eltanbouly, May Bashendy, and Tamer Elsayed. 2019. Simple but not naïve: Fine-grained Arabic dialect identification using only n-grams. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 214–218, Florence, Italy. Association for Computational Linguistics.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. DarijaBERT: a step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*.

Dhaou Ghoul and Gaël Lejeune. 2019. MICHAEL: Mining character-level patterns for Arabic dialect identification (MADAR challenge). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 229–233, Florence, Italy. Association for Computational Linguistics.

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

Nizar Habash, 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*. Morgan & Claypool Publishers.

Fei Huang. 2015. Improved Arabic dialect classification with social media data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural*

Language Processing, pages 2118–2126, Lisbon, Portugal. Association for Computational Linguistics.

Salma Jamal, Aly M .Kassem, Omar Mohamed, and Ali Ashraf. 2022. On the Arabic dialects' identification: Overcoming challenges of geographical similarities between Arabic dialects and imbalanced datasets. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 458–463, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nicola Jones. 2015. Artificial-intelligence institute launches free science search engine. *Nature*.

Vani Kanjirangat, Tanja Samardzic, Fabio Rinaldi, and Ljiljana Dolamic. 2022. Early guessing for dialect identification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6417–6426, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alan S. Kaye and Judith Rosenhouse. 1997. Arabic dialects and maltese. In Robert Hetzron, editor, *The Semitic Languages*, Routledge Language Family Series, pages 263–311. Routledge, London & New York.

Saméh Kchaou, Fethi Bougares, and Lamia Hadrich-Belguith. 2019. LIUM-MIRACL participation in the MADAR Arabic dialect identification shared task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 219–223, Florence, Italy. Association for Computational Linguistics.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. 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.

Amr Keleg, Walid Magdy, and Sharon Goldwater. 2024. Estimating the level of dialectness predicts inter-annotator agreement in multi-dialect Arabic datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 766–777, Bangkok, Thailand. Association for Computational Linguistics.

Abdullah Khered, Ingy Abdelhalim Abdelhalim, and Riza Batista-Navarro. 2022. Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 479–484, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ahmed El Kholy and Nizar Habash. 2012. Orthographic and morphological processing for English-Arabic statistical machine translation. *Machine Translation*, 26(1/2):25–45.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.

Javier A. Lopetegui, Arij Riabi, and Djamé Seddah. 2025. Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 168–181, Abu Dhabi, UAE. Association for Computational Linguistics.

Leena Lulu and Ashraf Elnagar. 2018. Automatic Arabic dialect classification using deep learning models. *Procedia Computer Science*, 142:262–269. Arabic Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Karen McNeil. 2018. *Tunisian Arabic Corpus: Creating a Written Corpus of an 'Unwritten' Language*, page 30–55. Edinburgh University Press.

Abir Messaoudi, Ahmed Cheikhrouhou, Hatem Haddad, Nourchene Ferchichi, Moez BenHajhmida, Abir Korched, Malek Naski, Faten Ghriss, and Amine Kerkeni. 2022. TunBERT: Pretrained contextualized text representation for Tunisian dialect. In *Intelligent Systems and Pattern Recognition*, pages 278–290, Cham. Springer International Publishing.

Mai Mohamed Eida, Mayar Nassar, and Jonathan Dunn. 2024. How well do tweets represent sub-dialects of Egyptian Arabic? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 41–55, Mexico City, Mexico. Association for Computational Linguistics.

Hamdy Mubarak. 2018. Dial2MSA: A tweets corpus for converting dialectal Arabic to Modern Standard Arabic. *OSACT*, 3:49.

Helene Olsen, Samia Touileb, and Erik Velldal. 2023. Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus. In *Proceedings of ArabicNLP 2023*, pages 370–384, Singapore (Hybrid). Association for Computational Linguistics.

Dilworth B. Parkinson. 1991. Searching for modern fusha: Real-life formal arabic. *al-'Arabiyya*, 24:31–64.

Pavel Přibáň and Stephen Taylor. 2019. ZCU-NLP at MADAR 2019: Recognizing Arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 208–213, Florence, Italy. Association for Computational Linguistics.

Ahmad Ragab, Haitham Seelawi, Mostafa Samir, Abdelrahman Mattar, Hesham Al-Bataineh, Mohammad Zaghloul, Ahmad Mustafa, Bashar Talafha, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2019. Mawdoo3 AI at MADAR shared task: Arabic fine-grained dialect identification with ensemble learning. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 244–248, Florence, Italy. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2268–2273, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. AL-QASIDA: Analyzing llm quality and accuracy systematically in dialectal Arabic. *Preprint*, arXiv:2412.04193.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube dialectal Arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland. European Language Resources Association (ELRA).

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 Sameer Salloum. 2018. *Machine Translation of Arabic Dialects*. PhD thesis, Columbia University.

Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Mohammed Attia, Mohamed Eldesouki, and Kareem Darwish. 2019. QC-GO submission for MADAR shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 290–294, Florence, Italy. Association for Computational Linguistics.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1239–1248, Valencia, Spain. Association for Computational Linguistics.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. ADI17: A fine-grained Arabic dialect identification dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein Al-Natsheh. 2020. Multi-dialect Arabic BERT for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118, Barcelona, Spain (Online). Association for Computational Linguistics.

Bashar Talafha, Ali Fadel, Mahmoud Al-Ayyoub, Yaser Jararweh, Mohammad AL-Smadi, and Patrick Juola. 2019. Team JUST at the MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 285–289, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Taha Tobaili. 2016. Arabizi identification in Twitter data. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 51–57, Berlin, Germany. Association for Computational Linguistics.

Samia Touileb. 2020. LTG-ST at NADI shared task 1: Arabic dialect identification using a stacking classifier. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 313–319, Barcelona, Spain (Online). Association for Computational Linguistics.

Jihen Younes, Hadhemi Achour, and Emna Souissi. 2015. Constructing linguistic resources for the tunisian dialect using textual user-generated contents on the social web. In *Current Trends in Web Engineering*, pages 3–14, Cham. Springer International Publishing.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. Language variety identification with true labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.

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 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

A Was Regional-level ADI Already Solved?

When framing a multi-label task as a single-label one, there is an expected maximal accuracy that an oracle model can achieve. For a sample with multiple valid labels, the gold-standard label and the prediction of the oracle model will both be randomly selected from the sample's set of valid labels. Both the randomly sampled gold standard label and the model's prediction should match for the prediction of the model to be considered correct. Keleg and Magdy (2023) introduced Equation 1 for estimating the expected maximal accuracy given the distribution of the number of labels in which a sentence is valid. Applying the formula to the regional-level labels of the 978 DA samples we used for our analysis, we get an expected maximal accuracy of 63.06% as per Equation 2. Such a low accuracy upper bound provides more evidence for modeling the task as a multi-label classification one.

$$\mathbf{E}[\mathbf{Accuracy_{max}(Dataset)}] = Perc_1 + \sum_{n=2}^{n=N_{dialects}} \frac{Perc_n}{n} \tag{1}$$

$$\mathbf{E}[\mathbf{Accuracy_{max}(NADI\ 2024_{regional})}] = 44 + \frac{18}{2} + \frac{14}{3} + \frac{12}{4} + \frac{12}{5} \approx 63.06\% \tag{2}$$

Test Set(s) Information and Label Distribution	**Results**
- AOC [*]: A random 10% of the dataset (>110K samples) *MSA* (>60% of the samples) - *EGY* - *LEV* - *GLF* (Zaidan and Callison-Burch, 2014)	Acc = 81%[†]
- AOC: *MSA* (6,355) - *EGY* (1,050) - *LEV* (1,050) - *GLF* (1,050) **- FB test set**: *MSA* (1,363) - *EGY* (800) - *LEV* (123) - *GLF* (96) (Huang, 2015)	Acc = 87.8%[†] Acc = 68.2%
VarDial 2016: *MSA* (274) - *EGY* (315) - *LEV* (344) - *GLF* (256) - *NOR* (351) (Malmasi et al., 2016)	Acc = 51.2%[†]
VarDial 2017: *MSA* (262) - *EGY* (302) - *LEV* (334) - *GLF* (250) - *NOR* (344) (Zampieri et al., 2017)	$F1_{weighted}$ = 0.763 [Sp]
VarDial 2018 (Broadcast): *MSA* (262) - *EGY* (302) - *LEV* (334) - *GLF* (250) - *NOR* (344) **+ VarDial 2018 (YouTube)**: *MSA* (944) - *EGY* (1,143) - *LEV* (1,131) - *GLF* (1,147) - *NOR* (980) (Zampieri et al., 2018)	$F1_{macro}$ = 0.589 [Sp]
MADAR (CORPUS-6): *MSA* (2,000) - *BEIRUT* (2,000) - *CAIRO* (2,000) - *DOHA* (2,000) - *TUNIS* (2,000) - *RABAT* (2,000) (Salameh et al., 2018)	Acc. = 93.6%[†]
Arabic Dialects Dataset: A subset of AOC and a Tunisian Corpus *EGY* (1,741) - *GLF* (1,092) - *LEV* (1,056) - *MSA* (1,600) - *NOR* (1,584) (El-Haj et al., 2018)	Acc = 66.12%[†]
Habibi [*]: A random 30% of the Habibi dataset (50,550 samples) Egyptian (27.7%) - Levantine (24.1%) - Gulf (18.3%) - Sudan (13.0%) - Iraqi (10.5%) - Meghribi (6.4%) (El-Haj, 2020)	Acc = 72.6%[†]

Table A1: The performance of regional-level ADI systems introduced in 8 different papers. The result of the best-performing model in each paper is reported. **Note**: [*]: the exact number of samples in each split is not explicitly reported and the used data splits could not be found, [†]: the train/test sets are based on random sampling from the same dataset (i.e., the same data distribution), [Sp]: the models' predictions are also based on additional speech features provided by the shared task organizers.

We contrasted the maximal estimated accuracy of 63.06% to the results of 8 different regional-level ADI papers, summarized in Table A1. Two issues arise in analyzing the results, which might have led to inflated models' performances. First, 5 papers used random train/test splits. Consequently, the test set's samples come from the same distribution as the training set, which was previously found to be problematic (Søgaard et al., 2021). Second, five papers reported accuracy scores on imbalanced test sets, for which macro-averaged F1-scores are more appropriate. Despite these two performance-inflating issues, all the reported scores still indicate that the task is not solved, except for the *MADAR (Corpus-6)* dataset (Salameh et al., 2018), for which we identify two potential reasons. MADAR's authors identified Beirut, Cairo, Doha, Tunis, and Rabat as anchor cities for wider regional dialects. Hence, sentences written in these city-level dialects might have been more distinguishable from each other compared to sentences from other non-anchor cities. Moreover, the dataset was created by translating the same sentences from English or French into MSA in addition to the 5 city dialects. The translators might have tried to include

more cues of their dialects in their translations to distinguish them from MSA translations and the other dialects' translations.

B Interannotator Agreement Scores for the Jordanian and Saudi Annotations

Country	Validity labels			ALDi ratings
	Fleiss κ	N valid	N ¬valid	Krip. α
Jordan	0.56	617 (455)	503 (367)	0.62
Saudi Arabia	0.62	476 (328)	644 (490)	0.65

Table B2: The Interannotator agreement scores for the validity labels and ALDi ratings, Fleiss' Kappa (κ) for Validity labels and Krippendorff's Alpha –interval method– (α) for ALDi ratings. *N valid* and *N ¬valid* represent the number of samples whose majority vote labels are *valid* and *not valid*, respectively, with the number of sentences with complete agreement reported (between brackets).

We extended the annotations of the 1,120 samples of the NADI 2024 by recruiting 3 annotators from Jordan and 3 from Saudi Arabia. The interannotator agreement scores are reported in Table B2. For the validity labels of each country, we compute the chance-corrected Fleiss' Kappa (κ) score, finding adequate agreement between the annotators of both countries. For the ALDi ratings, we use Krippendorff's Alpha –interval method– (α) between the numeric values of the ratings of each country's valid samples, which penalizes disagreements differently according to their assigned values. The range of the α scores is -1 to 1, with 0 indicating chance agreement. Hence, 0.62 and 0.65 signify that the annotators' agreement is substantially better than random, despite the subjectivity of the task.

C Country-level Overlap

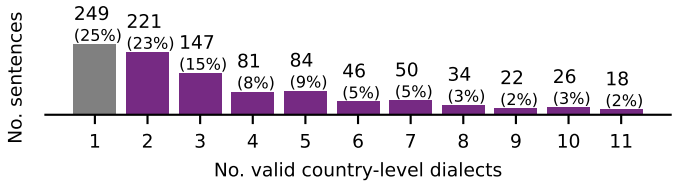


Figure C1: The histogram of the number of dialects in which a sentence is valid on the country-level dialects.

We compute the percentage of the samples within our dataset that are manually labeled as valid in multiple country-level dialects by annotators from these countries, to extend Abdul-Mageed et al.'s (2024) analysis by covering two additional country-level dialects. Only 249 sentences ($\approx$ 25%) are single-label as per Figure C1, compared to the $\approx$ 30% reported for 9 country-level dialects on NADI 2024's development set (Abdul-Mageed et al., 2024). This indicated that incorporating more country-level dialects would still increase the already high percentage of multi-label samples.

We also show the cross-country overlap in Figure C2. While it is clear that countries within the same region overlap more with each other, a substantial overlap with countries from other regions exists. Theoretically, our dataset is uniformly representative of the 14 different countries to which the samples were geolocated. However, the NADI 2024's authors found that the precision of their geolocation methodology varies for the different countries, and is the lowest for the countries of the Maghreb region (49.3% for Tunisia, 57.3% for Morocco, and 65.3% for Algeria). Hence, we think that further investigations are required before using these percentages as proxies for proximity between dialects.

D Lexical Cues

The TWT15DA is an ADI dataset built by iteratively augmenting lists of lexical cues of 15 country-level dialects using geolocated tweets having any of these cues, then streaming more geolocated tweets using the augmented lists (Althobaiti, 2022). For each country, the new cues to be added are non-MSA unigrams

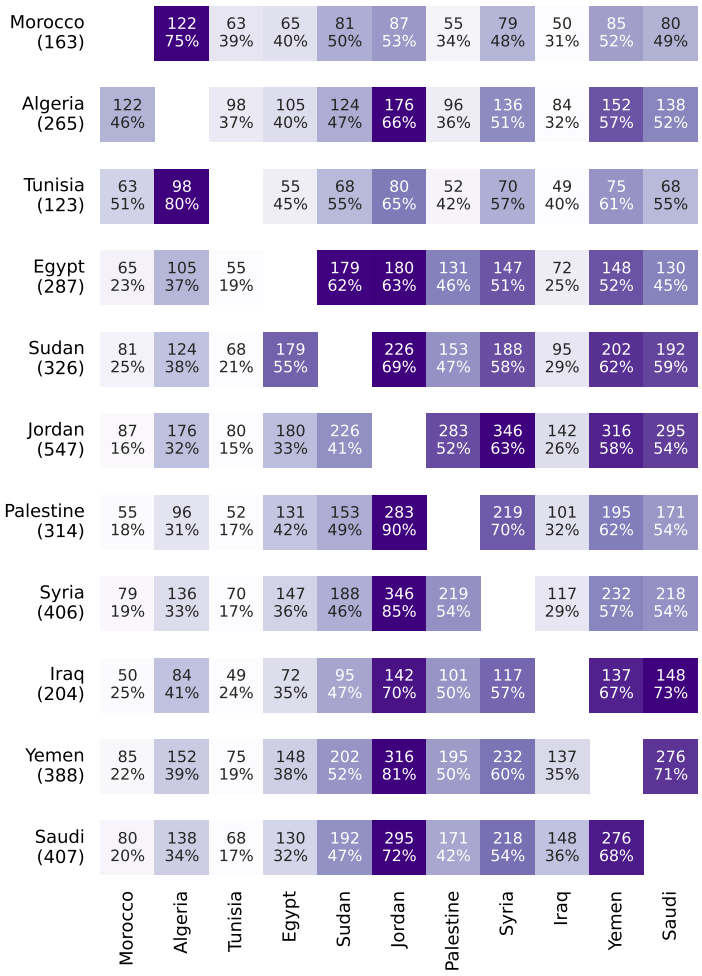


Figure C2: The percentage and number of each row country's valid samples that are also valid in the column country. **Note:** Each row's colormap range is independent from the other rows.

(a) in the tweets geolocated to this country, that **(b)** have high PMI values based on the following equation: $PMI(Unigram, Country) = log(\frac{P(Unigram, Country)}{P(Unigram)*P(Country)})$; where the probabilities are computed using maximum likelihood estimation. Therefore, the same unigram could have PMI scores for multiple countries (e.g., كيفاش /kyfAš/ in Algerian, Moroccan, and Tunisian Arabic lists with PMI scores of 2.07, 1.55, 1.19). Hence, these cues are not necessarily distinctive of a single country-level dialect. However, the author defines the *cues* as "words used in one or more Arabic dialects but never used in MSA, thereby distinguishing Arabic dialects from MSA".

We replicate the analysis in §4.3 for the TWT15DA dataset, and report the precision, recall, and distinctiveness scores in Table D3. Notably, the lists have a low range of precision scores $[0.31, 0.70]$, and an even lower range of distinctiveness scores $[0.02, 0.57]$.

Applying a Region's Lexical Cues only to the Region's Geolocated Samples For the TWT15DA dataset, each sample should have at least a cue for one of the dialects. However, the assigned label is based on the sample's geolocation, and not on the dialects associated to the cues. Hence, to assign a sample to a country-level dialect, the sample should **(a)** have a lexical cue of this dialect and **(b)** be geolocated to this country. To simulate this two-step method for each country's/region's list, we replicate our method, but then only consider the matching samples that are geolocated to the considered country/region. The results of applying this post-processing step for the three lists of cues (DART, DIAL2MSA, and TWT15DA) are reported in Table D4. The effectiveness of this step is better understood by contrasting the results in Table 1 and Table D3 to those in Table D4.

Country	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
Morocco	52	22	5	163	.42	.10	.13	410	45
Algeria	41	23	2	265	.56	.05	.09	421	38
Tunisia	62	19	1	123	.31	.02	.15	407	48
Egypt	33	23	18	287	.70	.55	.08	172	35
Jordan	51	30	1	547	.59	.02	.05	180	37
Syria	50	28	9	406	.56	.18	.07	94	28
Iraq	21	13	12	204	.62	.57	.06	179	18
Yemen	8	5	2	388	.62	.25	.01	137	8
Saudi	43	20	6	407	.47	.14	.05	145	26

Table D3: Lexical cues of the TWTDA15 datasets. **Note (1) :** For each region's list, we report the number of samples of our dataset matching any of the cues (M) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}), in addition to the total number of valid samples (N_{Val}). The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$). **Note (2):** The table lists the 9 countries that are common between the labels of our dataset, and the lists of TWT15DA which did not include *Palestine* and *Yemen*.

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	20	20	13	287	1.0	.65	.07	271	10
IRQ	6	6	6	204	1.0	1.0	.03	120	7
MGH	15	15	14	325	1.0	.93	.05	273	11
LEV	24	22	20	629	.92	.83	.03	240	8
GLF	0	0	0	407	-	-	.00	200	0

(a) DART's 5 regional lists.

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	25	25	15	287	1.0	.6	.09	28	10
MGH	34	32	29	325	.94	.85	.10	60	22
LEV	36	33	33	629	.92	.92	.05	31	11
GLF	0	0	0	407	-	-	.00	9	0

(b) DIAL2MSA's 4 regional lists.

Country	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
Morocco	15	14	5	163	.93	.33	.09	410	22
Algeria	13	13	2	265	1.0	.15	.05	421	18
Tunisia	12	9	1	123	.75	.08	.07	407	15
Egypt	13	13	11	287	1.0	.85	.05	172	14
Jordan	15	12	1	547	.80	.07	.02	180	14
Syria	12	11	5	406	.92	.42	.03	94	12
Iraq	11	11	11	204	1.0	1.0	.05	179	13
Yemen	4	4	2	388	1.0	.50	.01	137	6
Saudi	9	9	4	407	1.0	.44	.02	145	7

(c) TWT15DA's 9 country-level lists.

Table D4: The Precision (P), Distinctiveness (D), and Recall (R) of each region's/country's cues, when the matching samples not geolocated to the region/country are discarded. **Note:** For each region's list, we report the number of samples geolocated to this region, matching any of its cues (M) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}). The total number of samples valid in this regions (N_{Val}) are reported irrespective of their geolocations. The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$).

The range of the precision significantly improves to values > 0.9 for the three lists, except for the lists of Tunisia and Jordan in TWT15DA. The distinctiveness scores also improve, yet to much lower ranges compared to the precision. This hints that filtering out the samples that match any of a region's cues, yet are not geolocated to this region minimizes the impact of matching false friends of these cues, which are intuitively expected to be in samples geolocated to other regions.

Unsuprisingly, limiting the samples to ones geolocated to each list's region causes a decrease in the recall values, as all the samples valid in this region's dialect that are not geolocated to the region are pre-filtered. Another drawback of this geolocation-based step is that the samples' geolocations are not always available.