The Arabic Generality Score: Another Dimension of Modeling Arabic Dialectness

Sanad Shaban[1] **and Nizar Habash**[1,2]
[1]MBZUAI, [2]New York University Abu Dhabi
sanad.shaban@mbzuai.ac.ae, nizar.habash@nyu.edu

Abstract

Arabic dialects form a diverse continuum, yet NLP models often treat them as discrete categories. Recent work addresses this issue by modeling dialectness as a continuous variable, notably through the Arabic Level of Dialectness (ALDi). However, ALDi reduces complex variation to a single dimension. We propose a complementary measure: the Arabic Generality Score (AGS), which quantifies how widely a word is used across dialects. We introduce a pipeline that combines word alignment, etymology-aware edit distance, and smoothing to annotate a parallel corpus with word-level AGS. A regression model is then trained to predict AGS in context. Our approach outperforms strong baselines, including state-of-the-art dialect ID systems, on a multi-dialect benchmark. AGS offers a scalable, linguistically grounded way to model lexical generality, enriching representations of Arabic dialectness.[1]

1 Introduction

Arabic exhibits a well-known case of diglossia: Modern Standard Arabic (MSA) functions as the high variety in formal settings, while a range of Dialectal Arabic (DA) varieties are used in everyday communication (Ferguson, 1959). These dialects differ significantly from MSA, and from each other, in vocabulary, phonology, morphology, and syntax, often resulting in limited mutual intelligibility. And rather than forming discrete categories, they exist on a continuum. Token-level dissimilarity between MSA and dialects ranges from 37% to 67% (Salameh et al., 2018), and speakers frequently blend MSA and DA depending on context and background (Badawi, 1973; Badawi and Hinds, 1986; Iriarte Diez et al., 2023). Dialects themselves also share features and borrow from other languages, such as French and English (Hamed et al., 2020).

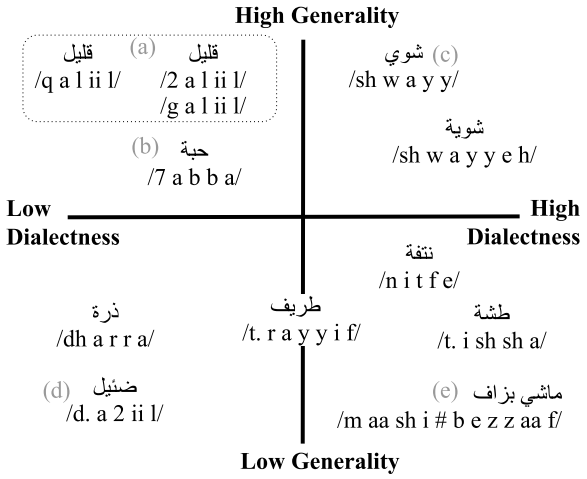

Figure 1: Different MSA and DA words and their CAPHI phonology representation (Habash et al., 2018), with varying generality and dialectness levels; all mean '(a) little (bit).'

Most computational models still frame dialect identification (DID) as a fixed-label classification task (Zaidan and Callison-Burch, 2014; Salameh et al., 2018; Abdul-Mageed et al., 2020, 2021, 2022, 2023), overlooking intra-dialect variation and cross-dialectal overlap. More recent approaches, like ALDi (Keleg et al., 2023), model dialectness as a continuous scalar, but reduce diverse signals to a single dimension.

We propose the Arabic Generality Score (AGS), a complementary dimension to ALDi that captures how broadly a word is used across dialects. Figure 1 illustrates how AGS enables distinctions between widely shared and highly localized dialectal items, even when both diverge from MSA. Together, ALDi and AGS define a two-dimensional space for modeling Arabic dialectness. We introduce a novel pipeline to annotate a parallel dialect corpus with word-level AGS, combining alignment, etymology- and phonology-aware edit distance, and smoothing. We then fine-tune a BERT-based regression model to predict AGS in context. Eval-

[1]Code is publicly available at https://github.com/CAMeL-Lab/arabic-generality-score

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30002–30013
November 4-9, 2025

uation on multi-dialect data shows that our model outperforms strong baselines and captures lexical generality more effectively.

Broader Impact We view AGS as a general-purpose signal with broad implications for dialectal Arabic NLP. While our primary contribution is the definition and estimation of AGS, the signal is directly relevant to several downstream tasks. (1) *Dialect Identification (DID):* AGS complements existing measures of dialectness (e.g., ALDi) by capturing lexical sharedness rather than only divergence from MSA, enabling finer-grained distinctions for ambiguous or unseen inputs. (2) *Dialect Rewriting and Controlled Generation:* By quantifying how regionally specific or pan-dialectal a lexical item is, AGS facilitates the substitution of localized expressions with more general alternatives, supporting applications in public service messaging, education, and inclusive media content. (3) *Evaluation and Benchmarks:* Prior work has shown that dialectness correlates with task difficulty (e.g., in translation; (Sajjad et al., 2020). AGS provides an orthogonal signal that enriches benchmarks and evaluations, offering a finer lens to assess model generalizability in the presence of dialectal variation (Kirchhoff et al., 2007).

The rest of the paper is organized as follows: Sections 2 and 3 review related work and relevant linguistic facts, respectively; Section 4 introduces AGS; Sections 5 and 6 detail the data and model; and Section 7 presents results.

2 Related Work

2.1 Arabic Dialect Identification

Arabic dialect identification (DID) has traditionally been treated as a single-label classification task, assigning each text a discrete dialect label such as Egyptian, Levantine, or Gulf Arabic, or Modern Standard Arabic (MSA) (Zaidan and Callison-Burch, 2014). Later work introduced finer-grained setups, including city-level classification with the MADAR-26 corpus (Bouamor et al., 2018; Salameh et al., 2018). However, such approaches assume dialectal boundaries are clean, despite evidence of overlap and mutual influence among varieties.

Recent work highlights the limitations of discrete labeling. Texts often exhibit features from multiple dialects and MSA, especially in social media. Error analyses show that many supposed misclassifications are linguistically plausible alternatives (Keleg and Magdy, 2023; Olsen et al., 2023), motivating multi-label and continuous approaches. The Arabic Level of Dialectness (ALDi) framework (Keleg et al., 2023), for instance, models dialectness on a continuous scale rather than as a hard class boundary. This shift is also evident in shared tasks. The NADI series evolved from strict classification in 2020 to multi-label and continuous subtasks by 2024 (Abdul-Mageed et al., 2020, 2021, 2024), including tasks for ALDi estimation.

Most recently, Keleg et al. (2025) demonstrated that some widely held simplifying assumptions in Arabic dialect research such as treating dialect identification as single-label, relying on fixed lexical cues, or ignoring cross-dialect annotation differences, can distort how datasets and models are constructed. They recommend multi-label DID formulations, more rigorous validation of lexical cues, and the use of ALDi scores to guide annotation and modeling.

These developments reflect a growing consensus: effective DID must account for the fluid, gradient nature of Arabic dialects, paving the way for multi-dimensional frameworks such as ours.

2.2 Dialectness

Dialectness refers to how much a text diverges from Modern Standard Arabic (MSA) and exhibits features of Dialectal Arabic (DA). Sociolinguistically, Badawi's five-level model (Badawi, 1973) captures this as a continuum from Classical Arabic to purely colloquial speech. It reflects how speakers shift registers based on context and education. Computationally, early approaches measured dialectness using word frequency ratios in DA versus MSA corpora (Zaidan and Callison-Burch, 2014; Sajjad et al., 2020). Later work introduced more detailed annotation schemes labeling tokens and segments by their deviation from MSA, incorporating orthographic, morphological, and lexical cues (Habash et al., 2008). While accurate, these methods are labor-intensive and require expert annotators.

Crowdsourced alternatives, such as the Arabic Online Commentary (AOC) corpus (Zaidan and Callison-Burch, 2011), used coarser labels (e.g., "Little" or "Mostly Dialectal") but lacked consistent guidelines. Building on these efforts, the Arabic Level of Dialectness (ALDi) (Keleg et al., 2023) introduced continuous sentence-level scores based on averaged crowd annotations. A regression model (Sentence-ALDi) fine-tuned on MarBERT predicts these scores on a 0 to 1 scale, where lower

values indicate MSA-like content and higher values correspond to strongly dialectal content. ALDi captures a fine-grained, replicable measure of dialectal divergence. However, it remains a single-axis metric and may conflate different types of variation, such as geographically localized versus widely used forms.

To address this limitation, we introduce the Arabic Generality Score (AGS), a complementary axis that captures how widely a word is used across dialects, independent of its divergence from MSA. AGS allows us to distinguish between regionally specific and broadly shared dialectal forms, information that ALDi alone cannot provide. Together, AGS and ALDi define a two-dimensional space of variation, offering a richer and more interpretable model of the Arabic dialect continuum.

3 Linguistic Background

3.1 Rich Morphology and Noisy Orthography

Arabic presents significant NLP challenges due to its rich morphology and highly inconsistent orthography, especially in dialects. Root-and-pattern morphology, combined with affixes and clitics, leads to high sparsity, while the lack of diacritics increases ambiguity. Dialects further complicate processing with informal spelling: a single word may appear in over 25+ forms (Eskander et al., 2013; Alhafni et al., 2024; Habash et al., 2018). In the absence of standardized conventions, spelling choices vary based on pronunciation and etymology. Writers may opt for phonetic spellings that reflect local speech or etymological ones that preserve MSA roots; sometimes these align, but often they diverge. For instance, consider the word for 'heart', قلب *qlb*[2] /q a l b/ in MSA. It is pronounced /2 a l b/ in Beirut, replacing the /q/ with a glottal stop /2/. A person from Beirut may choose to spell it phonetically as ألب *Âlb*, or etymologically, as قلب *qlb*.

3.2 CODA and CAPHI

We adopt a normalization framework designed to reduce surface variation and enhance cross-dialect comparability (Habash et al., 2018).

The Conventional Orthography for Dialectal Arabic (**CODA**) defines a consistent spelling system for dialects using the Arabic script. It reduces sparsity by mapping surface forms to their etymological roots where appropriate. In the example from Section 3.1, CODA maps ألب *Âlb* to قلب *qlb*, recognizing them as orthographic variants of the same underlying word.

The CAMeL Arabic Phonetic Inventory (**CAPHI**) provides a complementary phonological transcription layer specifically designed for Arabic. It captures dialect-specific pronunciations of CODA-normalized words. For instance, قلب *q-l-b* may be realized as /q a l b/, /2 a l b/, or /g a l b/ in MSA, BEI, and DOH, respectively.

Together, CODA and CAPHI expose structural equivalence across dialects and support normalization-aware approaches to dialect modeling.

4 The Arabic Generality Score

We introduce a new dimension of Arabic dialectness: the **Arabic Generality Score (AGS)** $\in [0, 1]$, which quantifies the extent to which an Arabic utterance is used across Arabic dialects and MSA. A higher AGS indicates broader generality across dialects and MSA, while a lower AGS reflects specificity. Although AGS may correlate with ALDi in parts of the dialect-MSA space, particularly where dialects intersect with MSA, it captures a distinct phenomenon. Many features with high AGS are not necessarily close to MSA. For example, dialectal terms such as ليش *lyš* 'why', مافي *mAfy* 'there isn't', and شوي *šwy*∼ 'a little' are widespread across Levantine, Gulf, and North African varieties. These items are perceived as highly *general* due to their cross-dialectal prevalence, despite their non-standard status. Conversely, expressions like السلام عليكم *AlslAm ʕlykm* 'Hello/Peace Be Upon You' are standard but also frequent in dialectal contexts. As illustrated in Figure 1, words span the full space defined by AGS and ALDi.

In the next two sections, we introduce the data used in this study and present a pipeline for computing and estimating word-level AGS using a parallel dialectal corpus.

5 Data and Resources

MADAR Corpus & MADAR-CODA The MADAR corpus (Bouamor et al., 2018) is a 26-way parallel dataset of 2,000 sentences from the Basic Travel Expression Corpus (BTEC), translated into MSA and 25 Arabic city dialects (**MADAR-26**). A subset of five dialects (CAI, BEI, DOH, TUN, and RAB) was extended with 10,000 additional

[2]HSB Arabic transliteration (Habash et al., 2007).

sentences each (**MADAR-6**). **MADAR-CODA** (Eryani et al., 2020) provides CODA-normalized versions of 2,000 MADAR-6 sentences.

MADAR Lexicon was also developed as part of the MADAR project, as a multilingual, multi-dialectal lexical resource, covering 1,045 concepts across 25 Arabic city dialects, along with English, French, and Modern Standard Arabic (MSA) (Bouamor et al., 2018). Each concept is defined using a triplet of words (En, Fr, MSA) and populated with dialectal variants annotated for both CODA orthography and CAPHI phonology. The lexicon includes 47,466 dialectal entries.

CAPHI Table The CAMeL Arabic Phonetic Inventory (CAPHI) table, part of the CODA* guidelines, supports consistent, phonologically informed orthographic choices across dialects (Habash et al., 2018). It includes: (1) a CAPHI column with phonemes, (2) a CODA column with their script representations, and (3) a default mapping. For example, the CAPHI phoneme /p/ is mapped to ب *b* in CODA, as in بري *bry* /p r i/ 'price' (Algiers), but is assigned the default phonetic value /b/. The table draws from a large number of dialects, guiding the standardized representation of sounds absent in MSA.

6 Methodology

Our methodology consists of five main components: aligning dialectal word pairs, computing a phonologically informed edit distance, aggregating these distances into a lexical similarity score, estimating this score for unseen words in context, and extending the measure from the word to the sentence level.

6.1 Word Alignment

The task of word alignment over parallel corpora refers to identifying semantically equivalent words across sentences in different dialects or languages. More formally, given a set of k parallel sentences

$$\mathbf{s} = \{s^{(d_1)}, s^{(d_2)}, \ldots, s^{(d_k)}\},$$

where each sentence is associated with a dialect label $d_i \in \mathcal{D}$ (e.g., MSA, CAI, BEI, etc.), and

$$s^{(d)} = \langle w_1^{(d)}, w_2^{(d)}, \ldots, w_{n_d}^{(d)} \rangle$$

is a tokenized sequence in dialect d, the goal is to extract sets of word correspondences that capture equivalent meaning:

$$\mathcal{A} \subseteq \left\{ \{w_p^{(d_i)}, w_q^{(d_j)}\} \;\middle|\; w_p^{(d_i)} \approx w_q^{(d_j)},\ d_i \neq d_j \right\},$$

where $\approx$ denotes semantic equivalence between words across dialects.

We use **AWESOME Align** (Dou and Neubig, 2021), a neural word alignment method based on multilingual contextual embeddings (see Appendix A). To handle multiple parallel sentences, we align each dialectal sentence to a central MSA anchor, assuming dialectal words aligned to the same MSA word are mutually aligned. See example in Table 1. More formally, let $d_1 = \text{MSA}$, and define:

$$\mathcal{A}_{\text{MSA}} = \left\{ \{w^{(d_1)}, w^{(d_2)}, \ldots, w^{(d_k)}\} \;\middle|\; w^{(d_i)} \approx w^{(d_1)},\ \forall i \in [2, k] \right\} \quad (1)$$

We aggregate alignments for each word–dialect pair across the entire corpus. Given a word w in dialect d, we define:

$$\mathcal{A}(w, d) = \left\{ a \in \mathcal{A}_{\text{MSA}} \;\middle|\; w^{(d)} \in a \right\},$$

where each $a \in \mathcal{A}_{\text{MSA}}$ is a set of semantically aligned words across dialects. This aggregation allows us to collect more occurrences of each word in cross-dialectal contexts, refining the signal used in our AGS modeling.

For example, consider the MSA word أردت *Ârdt* 'I/you wanted'. After aggregating all alignments of this word across MADAR-6, we obtain $\mathcal{A}(\text{أردت}, \text{MSA})$ illustrated in Table 2. Instead of just one reference, aggregation allows for multiple equivalent references in the other dialects.

Moving forward, we will use the aggregated alignments $\mathcal{A}(w, d)$ to compute the AGS for a word w in dialect d. Linking this back to the previous example, we will first have to compute the distance between the word أردت and all its dialectal counterparts. For that, we need to define an edit distance function that is robust under the lack of standard orthography in DA.

6.2 Augmented Edit Distance

Consider the edit distance between قلب *qlb* in DOH and ألب *Âlb* in BEI *'heart'*. The Levenshtein edit distance penalizes the first character

	MSA	CAI	BEI	DOH	RAB	TUN	English Gloss
(1)	إنها	هو	هوي	موجود	كاين	موجود	it is/ it exists
	ĂnhA	*hw*	*hwy*	*mwjwd*	*kAyn*	*mwjwd*	
(2a)	في	في	بأخر	في	في	في	(in) (the) end (of)
	fy	*fy*	*bÂxr*	*fy*	*fy*	*fy*	
(2b)	أخر	اخر	بأخر	نهاية	اللاخر	اخر	
	Âxr	*Axr*	*bÂxr*	*nhAyħ*	*AllAxr*	*Axr*	
(3)	القاعة	القاعة	الصالة	الممر	القاعة	الكولوار	corridor / hallway
	AlqAçħ	*AlqAçħ*	*AlSAlħ*	*Almmr*	*AlqAçħ*	*AlkwlwAr*	
(4a)	سوف	حأجيبلك	رح	بجيب	—	و	I will bring you
	swf	*HÂjyblk*	*rH*	*bjyb*	—	*w*	
(4b)	آتي	حأجيبلك	جبلك	بجيب	انجيب	نجيبهولك	
	Āty	*HÂjyblk*	*jblk*	*bjyb*	*Anjyb*	*njybhwlk*	
(4c)	لك	حأجيبلك	جبلك	لك	ليك	نجيبهولك	
	lk	*HÂjyblk*	*jblk*	*lk*	*lyk*	*njybhwlk*	
(5)	ببعض	شوية	شوي	شوي	شويا	—	some
	bbçD	*šwyħ*	*šwy*	*šwy*	*šwyA*	—	

Table 1: Word-level alignments between an MSA sentence and its dialectal equivalents across five Arabic varieties.

Dialect	Aligned Forms with Frequencies
MSA	أردت *Ârdt*
BEI	بدك *bdk* (4), بتحب *btHb* (1), بدي *bdy* (2)
CAI	محتاج *mHtAj* (1), عايز *çAyz* (5), عاوز *çAwz* (1)
TUN	تستحق *tstHq* (1), تحب *tHb* (3), نحب *nHb* (2), *None* (1)
DOH	احتجت *AHtjt* (1), بغيت *bγyt* (1), ابغي *Abγy* (2), بتوقف *btwqf* (1), تبغي *tbγy* (2)
RAB	حتاجيتي *HtAjyty* (1), بغييتي *bγyyty* (1), نبغي *nbγy* (1), بغيتي *bγyty* (1), بغيت *bγyt* (1), كنت *knt* (1), *None* (1)

Table 2: Aggregated alignments for the MSA word أردت *Ârdt* 'I want' across five dialects. Numbers in parentheses indicate frequency of occurrence in the corpus.

CODA	CAPHI	Dialect	CODA-CAPHI Alignment
جلد *jld*	/dj i l i d/	KHA	[(ج, dj),(-1, i),(ل, l),(-1, i),(د, d)]
جلد *jld*	/g e l d/	CAI	[(ج, g),(-1, e),(ل, l),(د, d)]
جلد *jld*	/j a l d/	RAB	[(ج, j),(-1, a),(ل, l),(د, d)]
جلد *jld*	/j i l i d/	BEI	[(ج, j),(-1, i),(ل, l),(-1, i), (د, d)]
جلد *jld*	/y i l d/	DOH	[(ج, y),(-1, i),(ل, l),(د, d)]

Table 3: Dialectal phonological variants of جلد *jld* 'leather' from the MADAR Lexicon, with letter-to-phoneme alignments.

substitution (ق q ↔ أ Â), though both forms share the same etymological root ق q. This variation stems from dialect-specific phonological realization (/g/ in DOH and /2/ in BEI) rather than true lexical divergence. To address this, we augment the Levenshtein algorithm with an etymology-aware substitution cost.

Formally, for any character, we can define three variables representing orthography, etymology and phonology:

$$(x_{or}, x_{et}, x_{ph}) = \textbf{(أ Â , ق } \boldsymbol{q}\textbf{, /2/)}$$

$$(y_{or}, y_{et}, y_{ph}) = \textbf{(ق } \boldsymbol{q}\textbf{ , ق } \boldsymbol{q}\textbf{, /g/)}$$

where only x_{or}, y_{or} are observed, and the underlying (x_{et}, y_{et}) are equivalent. To handle such cases, we define the substitution cost as:

$$\text{cost}(x_{\text{or}}, y_{\text{or}} \mid d_x, d_y) = 1 - P(x_{\text{et}} = y_{\text{et}} \mid x_{\text{or}}, y_{\text{or}}, d_x, d_y) \quad (2)$$

The substitution cost of x_{or} with y_{or} in dialects (d_x, d_y) is proportional to the probability they differ etymologically. Estimating this requires three components:

(x_{et}, x_{ph})	(Phon = Etym)	Count	Etym#
(ج *j*, /dj/)	1	1	1
(ج *j*, /g/)	0	1	1
(ج *j*, /j/)	1	2	2
(ج *j*, /y/)	0	1	1
(ل *l*, /l/)	1	5	0
(د *d*, /d/)	0	5	0

Table 4: Etymological mappings inferred from CODA-PHON alignments in Table 3.

1. Phoneme from Etymology and Dialect: $P(x_{\text{ph}} \mid x_{\text{et}}, d_x)$. This represents the probability of a phoneme x_{ph} given its etymological character x_{et} in dialect d_x. To estimate this distribution, we use the **MADAR Lexicon** (Bouamor et al., 2018). We align CODA characters to CAPHI phonemes via Levenshtein alignment, guided by the CAPHI mapping table (Habash et al., 2018), then compute the conditional probability as:

$$P(x_{\text{ph}} \mid x_{\text{et}}, d_x) = \frac{\text{count}(x_{\text{ph}}, x_{\text{et}}, d_x)}{\text{count}(x_{\text{et}}, d_x)}$$

2. Phoneme from Orthography and Dialect: $P(x_{\text{ph}} \mid x_{\text{or}}, d_x)$ is the probability of a phoneme given an orthographic character in a dialect. To estimate it, we extend the **MADAR Lexicon** with unnormalized forms from **MADAR-CODA**, aligning these to CAPHI phonemes to compute:

$$P(x_{\text{ph}} \mid x_{\text{or}}, d_x) = \frac{\text{count}(x_{\text{ph}}, x_{\text{or}}, d_x)}{\text{count}(x_{\text{or}}, d_x)}$$

3. Phoneme-Based Etymology Detection: The orthographic character x_{or} can sometimes directly reflect the etymological character x_{et}, particularly when the spelling is etymologically motivated. Certain grapheme-to-phoneme mappings, such as ق *q*→ /2/, may serve as indicators that $x_{or} = x_{et}$. We infer these etymologically consistent spellings from the **MADAR Lexicon** using the heuristic: *If a character within one CODA word maps to different default and non-default CAPHI phonemes across dialects, then the orthographic form must preserve the etymology.* An illustrative example from the MADAR Lexicon is presented in Tables 3 and 4. As ج j in the CODA word جلد *jld* was mapped to different default and non-default phonemes, we deduce that these mappings were etymological. We apply this rule to the MADAR Lexicon to estimate $P(x_{et} = x_{or} | x_{or}, x_{ph}, d_x)$.

Now that all components have been defined, we can compute the posterior probability of the etymological form x_{et} given the observed orthographic character x_{or} in dialect d_x. Using the law of total probability over possible phonemic realizations $x_{\text{ph}} \in$ CAPHI, we obtain:

$$P(x_{\text{et}} \mid x_{\text{or}}, d_x) = \sum_{x_{\text{ph}} \in \text{CAPHI}} P(x_{\text{et}} \mid x_{\text{or}}, x_{\text{ph}}, d_x) \cdot P(x_{\text{ph}} \mid x_{\text{or}}, d_x)$$

where the final term $P(x_{\text{ph}} \mid x_{\text{or}}, d_x)$ refers to the second component discussed above. To compute the first term, we condition on etymological spelling ($x_{or} = x_{et}$):

$$\begin{aligned} &P(x_{\text{et}} \mid x_{\text{or}}, x_{\text{ph}}, d_x) = \\ &P(x_{\text{et}} \mid x_{\text{or}}, x_{\text{ph}}, d_x, x_{et} = x_{or}) \cdot P(x_{et} = x_{or} \mid x_{\text{or}}, x_{\text{ph}}, d_x) \\ &\qquad + \\ &P(x_{\text{et}} \mid x_{\text{or}}, x_{\text{ph}}, d_x, x_{et} \neq x_{or}) \cdot P(x_{et} \neq x_{or} \mid x_{\text{or}}, x_{\text{ph}}, d_x) \end{aligned}$$

Since non-etymological spellings follow an etymology-to-phonology mapping, this reduces to:

$$\begin{aligned} &P(x_{\text{et}} \mid x_{\text{or}}, x_{\text{ph}}, d_x) = \\ &\mathbb{1}(x_{\text{et}} = x_{\text{or}}) \cdot P(x_{et} = x_{or} \mid x_{\text{or}}, x_{\text{ph}}, d_x) \\ &+ P(x_{\text{et}} \mid x_{\text{ph}}, d_x) \cdot P(x_{et} \neq x_{or} \mid x_{\text{or}}, x_{\text{ph}}, d_x) \end{aligned}$$

which we have prepared in the above-mentioned components 1 and 3. Finally, as defined before, the substitution cost between two characters x and y from dialects d_x and d_y as the probability that they *do not* share the same etymological origin (2). To estimate this probability, we sum over all latent etymological character c (i.e., possible CODA representations):

$$\begin{aligned} &P(x_{\text{et}} = y_{\text{et}} \mid x, y, d_x, d_y) = \\ &\sum_{c \in \text{CODA}} P(x_{\text{et}} = c \mid x_{\text{or}}, d_x) \cdot P(y_{\text{et}} = c \mid y_{\text{or}}, d_y) \end{aligned} \quad (3)$$

Each of the individual terms is computed as described above, by marginalizing over phonemic realizations and conditioning on whether the orthographic form is etymological. Furthermore, we can retrieve the most probable etymology for the substitution by identifying the tuples $(x_{\text{et}}, x_{\text{or}}, x_{\text{ph}})$ and $(y_{\text{et}}, y_{\text{or}}, y_{\text{ph}})$ that maximize the product in equation (3). These tuples correspond to the most likely alignment path through the etymological space.

6.3 Aggregating Distances into AGS

We now aggregate the distances into a scalar AGS. Intuitively, a word is more "general" if it closely aligns with words from many other dialects. For each dialect, we select the *minimum* distance between w and any of its aligned forms from that dialect. This yields a dictionary of minimum distances $\{\delta_d\}$, where each δ_d represents the closest aligned counterpart from dialect d.

High distance values often reflect coincidental overlap rather than true etymological similarity, for instance, a distance of 0.8 may be no more meaningful than 1.0. To address this, we apply a smoothed threshold using a logistic function, which softly weights distances based on proximity to the cutoff, rather than applying a hard filter:

$$f(d) = \frac{1}{1 + e^{-s(t-d)}} \tag{4}$$

The function outputs values near 1 for distances well below the threshold t, and near 0 for those far above. It is symmetric around $f(t) = 0.5$, with s controlling steepness. This enables smooth weighting of alignment quality without hard cutoffs.

The final AGS is computed by averaging the smoothed values across dialects:

$$\text{AGS}(w) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} f(\delta_d) \tag{5}$$

We test three threshold settings with fixed steepness $s = 20$:

$$(t, s) \in \{(0.3, 20),\ (0.4, 20),\ (0.5, 20)\}$$

Lower t values enforce stricter alignment, while higher ones permit looser matches. Fixing s ensures a sharp transition near t.

6.4 Estimating AGS

In the final stage, we train a regression model to predict the AGS of a word in context. Given a word w from dialect d, we retrieve one or more sentences from the MADAR corpus in which w appears. Each sentence is pre-processed by marking the target word using dedicated special tokens. For example, if the sentence is: أنا متوقف عن العمل *ÂnA mtwqf ς*n *Alςml* and the target word is متوقف *mtwqf*, we transform the sentence to:

أنا [TGT] متوقف [TGT] عن العمل

*ÂnA [TGT] mtwqf [TGT] ς*n *Alςml*.

These special tokens help the model focus on the word of interest and learn a mapping from its context to the AGS. We fine-tune the pretrained **CAMeL-BERT**[3] model (Inoue et al., 2021) for regression, using mean squared error (MSE) loss.

6.5 From Word-Level to Sentence-Level AGS

To extend our word-level AGS to the sentence level, we propose an aggregation method that accounts for the disproportionate impact of specific words. Intuitively, the presence of just one highly specific word can drastically reduce the perceived generality of an entire sentence.

We compute the harmonic mean over the k lowest-scoring words in each sentence. The harmonic mean penalizes low values more heavily than the arithmetic mean, making it well-suited for capturing the influence of specific items on overall sentence AGS. Formally, let $s = \{w_1, w_2, \ldots, w_n\}$ be a sentence with $g(w_1), g(w_2), \ldots, g(w_n)$ being word-level AGS's. Let $\{g_1, g_2, \ldots, g_k\}$ be the k lowest scores in the sentence. The sentence-level AGS $G(s)$ is computed as:

$$G(s) = \frac{k}{\sum_{i=1}^{k} \frac{1}{g_i}} \tag{6}$$

The choice of k is treated as a tunable hyperparameter, optimized in downstream experiments.

7 Experiments and Results

7.1 AGS Statistics

Figure 2 presents the distribution of AGS annotations, categorized as Specific (0-0.1), Moderate (0.1-0.5), and General (0.5-1.0) across six varieties in MADAR-6. MSA exhibits the highest proportion of Specific words (∼39%). In contrast, DOH (Doha) and BEI (Beirut) show a strong skew toward General words, with DOH reaching over 43%, suggesting broader lexical overlap across dialects. Moderate AGS scores are more evenly distributed, with all dialects clustering around 30–34%. The trend highlights dialectal variation in lexical generality and suggests that some dialects (e.g., DOH, BEI) may serve as better hubs for cross-dialectal generalization. This relatively low generality in MSA may be tied to the stylistic and structural nature of its translations. MSA forms are often longer,

[3] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix

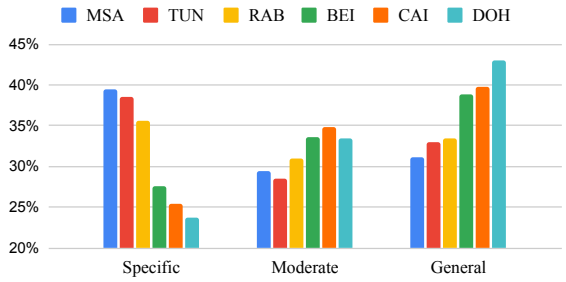

Figure 2: Distribution of Specific (0-0.1), Moderate (0.1-0.5), and General (0.5-1) annotations in AGS-annotated MADAR-6.

Model	RMSE
Trained Models	
CAMeL-BERT on MADAR-26	**0.2698**
CAMeL-BERT on MADAR-6	0.2704
Baselines	
X1: MADAR Lookup	0.2901
X2: B2BERT	0.3003
X3: NADI2024-baseline	0.2985

Table 5: RMSE for AGS estimation across trained models and baselines on MDID-test.

more formal, and semantically over-specified compared to dialectal variants. Concretely, MSA translations have the highest average sentence length in both characters (35.8) and words (8.0), compared to dialects like DOH (27.8 characters, 5.3 words) and BEI (28.5 characters, 5.6 words). These inflated constructions reduce surface-level overlap with other varieties.

7.2 Evaluation of AGS Estimation

Test Set For evaluation, we use the MDID-DEV and MDID-TEST datasets from the NADI 2024 shared task (Abdul-Mageed et al., 2024), which include 120 and 1000 Arabic tweets, respectively, annotated for dialectal validity across 11 country-level dialects. We convert the multi-label annotations into a scalar *sentence-level AGS* by computing the ratio of valid dialect labels to the total number of dialects:

$$\text{AGS}_{\text{sent}} = \frac{n_{\text{valid}}}{n} \tag{7}$$

where n_{valid} is the number of dialects a sentence is annotated with, and n is the total dialect count. This matches our AGS definition: broader validity implies more general vocabulary. We then apply our *word-level AGS model* to each sentence by averaging predicted AGS scores across its words (Section 6.4).

Metric We evaluate the predicted sentence-level scores against the ground truth AGS derived from MDID using the **Root Mean Squared Error (RMSE)** $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ is the predicted AGS and y_i is the ground truth score for sentence i.

Baselines We compare our method against three baselines: (X1) **MADAR Lookup**, which assigns AGS values based on a direct lookup from the annotated MADAR-26 lexicon, defaulting to 0.5 for out-of-vocabulary words; (X2) **B2BERT**, from the NADI 2024 leaderboard, which uses binary classifiers per dialect and achieves 0.5963 macro-F1 on the MDID test set;[4] and (X3) the **NADI2024-baseline**, an official baseline using a top-p inference strategy, which achieves 0.4697 macro-F1.[5] For both X2 and X3, dialect predictions are converted to sentence-level AGS by averaging over predicted dialect labels, using our aggregation method (Section 6.5).

Discussion Table 5 shows that both of our AGS models, **CAMeL-BERT on MADAR-6** and **CAMeL-BERT on MADAR-26**, outperform all baselines, with the latter achieving the lowest RMSE. This suggests that finetuning specifically for AGS yields better generality estimates than MDID-based models. The small performance gap between MADAR-6 and MADAR-26 suggests that a small geographically-diverse set of dialects (as in MADAR-6) has a strong generality signal relative to a more fine-grained 26-variety setup.

Table 6 illustrates how AGS captures generality in context. The sentences from the MDID dataset were annotated as valid in one and two out of 11 dialects, corresponding to sentence-level AGS scores of 0.091 and 0.273, respectively. In the first example, high-AGS words such as من *mn* and مين *myn* are broadly used across dialects, while lower-scoring terms like مخنوق *mxnwq* and طايق *TAyq* are more dialect-specific. In the second example, مافي *mAfy* scores highly due to its wide regional usage, whereas expressions like واخرتنا *wAxrtnA* and مفر *mfr* receive lower scores. The two least general

[4] https://huggingface.co/AHAAM/B2BERT
[5] https://huggingface.co/AMR-KELEG/NADI2024-baseline

AGS: 0.091 , **Predicted AGS:** 0.173

Word	Gloss	AGS
مين *myn*	who	0.986
هيكون *hykwn*	will be	0.725
جنبك *jnbk*	beside you	0.817
ف *f*	in	0.941
حزنك *Hznk*	your sadness	0.549
وزعلك *wzʕlk*	and your upset	0.448
مين *myn*	who	0.986
هيكون *hykwn*	will be	0.725
جنبك *jnbk*	beside you	0.817
وانت *wAnt*	while you are	0.637
مخنوق *mxnwq*	suffocating	0.229
ومش *wmš*	and not	0.371
طايق *TAyq*	tolerating	0.139
الدنيا *AldnyA*	the world	0.498

AGS: 0.273 , **Predicted AGS:** 0.278

Word	Gloss	Score
مافي *mAfy*	there is no	0.821
مفر *mfr*	escape	0.312
من *mn*	from	0.987
هالشغله *hAlšglh*	this thing	0.525
واخرتنا *wAxrtnA*	and eventually	0.251
بدنا *bdnA*	we want	0.465
نكمل *nkml*	to complement	0.812
نص *nS*	half	0.943
ديننا *dynnA*	our religion	0.431

Table 6: Two MDID sentences with word-level AGS predicted by CAMeL-BERT (MADAR-26). Sentence-level AGS is computed as the geometric mean (eq. 6) of the two least-general (k=2) highlighted words.

words in each case drive the predicted AGS.

8 Conclusions and Future Work

We introduced the Arabic Generality Score (AGS), a new measure of how widely a word is used across dialects, complementing existing metrics like ALDi. We built an annotation pipeline and trained a contextual model that outperforms strong baselines on a multi-dialect benchmark.

Future work includes extending AGS to phrases and constructions, modeling variation across domains and time, and integrating AGS into downstream tasks like translation, retrieval, and educational NLP.

Limitations

While AGS offers a promising step toward multi-dimensional modeling of Arabic dialectness, it has several limitations. First, our method depends on the availability of parallel dialectal corpora, which are still scarce and often skewed toward certain regions such as the Levant or Egypt. This limits the diversity of dialectal features in the data.

Second, the annotation pipeline assumes that surface similarity aligns with functional or semantic equivalence across dialects, which is not always the case, particularly with idioms or culturally specific expressions.

Third, the model may overfit to frequent patterns and underperform on low-resource varieties. We also do not currently include geographic or sociolinguistic metadata, which may help improve AGS estimation.

Ethics Statement

Our work aims to support more inclusive, dialect-aware Arabic NLP. However, dialect modeling can raise concerns related to identity, marginalization, and language hierarchies. AGS is a descriptive tool, not a value judgment of language use.

We used only publicly available corpora and followed standard ethical practices. Still, dialectal data may reflect existing biases, including under-representation of certain regions or groups. We encourage future efforts to broaden dialect coverage.

Finally, AGS models could be misused for profiling or surveillance. We strongly oppose any application of this work in ways that compromise privacy or reinforce discrimination.

We used AI writing assistance within the scope of "Assistance purely with the language of the paper" described in the ACL Policy on Publication Ethics.

Acknowledgments

We thank Amr Keleg for making his research and resources publicly available, which our work builds upon, and for generously providing access to the MDID n_label data used in our evaluation.

References

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. 2020. 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.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

El-Said M. Badawi and Martin Hinds. 1986. *A Dictionary of Egyptian Arabic*. Librairie du Liban, Beirut.

Said M. Badawi. 1973. *Mustawayāt al-çarabiyya al-muāşira fī Miṣr*. Dār al-Maçārif bi-Miṣr, Cairo.

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

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

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

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

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

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

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

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ana Iriarte Diez, Claudia Laaber, Nina Kampen, and Monserrat Fernández. 2023. What is white arabic? new labels in a changing arab world. *Revista Española de Lingüística*, 53:229–266.

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, Sharon Goldwater, and Walid Magdy. 2025. Revisiting common assumptions about Arabic dialects in NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327, Vienna, Austria. 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.

Katrin Kirchhoff, Owen Rambow, Nizar Habash, and Mona Diab. 2007. Semi-automatic error analysis for large-scale statistical machine translation. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

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.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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.

A AWESOME Align

Given a pair of sequences

$$x = \langle x_1, \ldots, x_n \rangle \quad \text{and} \quad y = \langle y_1, \ldots, y_m \rangle,$$

AWESOME (Dou and Neubig, 2021) extracts contextual embeddings

$$h_x \in \mathbb{R}^{n \times d}, \quad h_y \in \mathbb{R}^{m \times d},$$

from a shared encoder and computes a similarity matrix:

$$S = h_x h_y^\top.$$

A row-wise softmax yields the alignment probability matrix:

$$S_{xy}^{(i,j)} = \frac{\exp(h_{x_i} \cdot h_{y_j})}{\sum_{j'=1}^{m} \exp(h_{x_i} \cdot h_{y_{j'}})}.$$

Alignments are extracted by applying a symmetric agreement criterion: a word pair (x_i, y_j) is aligned if the forward and backward probabilities exceed a confidence threshold τ:

$$(x_i, y_j) \in \mathcal{A} \quad \text{iff} \quad S_{xy}^{(i,j)} > \tau \quad \text{and} \quad S_{yx}^{(j,i)} > \tau.$$

We finetuned the encoder CAMeLBERT (Inoue et al., 2021) using all of AWESOME's proposed objectives: (1) Masked Language Modeling (MLM) , applied independently on source and target sentences, (2) Translation Language Modeling (TLM) , MLM on concatenated source–target pairs to encourage cross-sentence alignment, (3) Self-Training Objective (SO) , promotes closeness of initially aligned token embeddings, (4) Parallel Sentence Identification (PSI) , contrastive loss distinguishing parallel from non-parallel sentences, and (5) Consistency Optimization (CO) , maximizes agreement between forward and backward alignment matrices. We finetuned CAMeLBERT on approximately 100,000 MSA–DA sentence pairs from the MADAR-26 corpus, covering 25 dialects. Although MADAR contains no gold-standard alignments, previous work has shown that these objectives significantly improve alignment accuracy.

B Implementation

Hyperparameters For the aggregation function (4), we evaluated three threshold values $t \in \{0.3, 0.4, 0.5\}$ and selected $t = 0.5$ based on the lowest RMSE for AGS estimation on the MDID-DEV set. For contextual AGS prediction, we fine-tune CAMeL-BERT using a batch size of 32, a learning rate of 4×10^{-5} with a linear decay schedule and zero warmup steps, and the AdamW optimizer. Training was capped at 10,000 steps with early stopping based on validation performance. All experiments were run with a fixed random seed of 42.

Hardware All experiments were conducted on Google Colab using an NVIDIA A100 GPU. Training checkpoints were saved every 1,000 steps and monitored using Weights & Biases.

Libraries and Tools We used Python with `pandas`, `numpy`, and `camel-tools` for data preprocessing and orthographic normalization. Sentence alignment relied on `AWESOME-Align`, and contextual modeling was implemented with `transformers (v4.x)` and `PyTorch`.