A Corpus-Based Investigation of Contemporary Arabic Dialects Using the SADA Corpus

Ghada Alfattni
Computer Science Department
Jamoum University College
Umm Alqura University
gafattni@uqu.edu.sa

Abstract

The spoken Arabic exhibits substantial dialectal variation in the Arabic-speaking world. This paper presents a corpus-based analysis of Arabic dialectal variation using the SADA corpus, examining lexical, morphosyntactic, and discourse-pragmatic patterns across dialects. We combine quantitative frequency-based measures with qualitative linguistic analysis, including keyword comparison, distributional profiling, collocational and trigram analyses, and similarity-based clustering. Our results show that Arabic dialects share a substantial common core, while differing systematically in frequent discourse markers, evaluative expressions, and recurrent phraseological patterns. These findings provide empirical evidence for regional clustering among contemporary dialects and for variation relative to the standard register. The study contributes linguistic insights that support both Arabic dialectology and the development of dialect-aware NLP systems.

1 Introduction

Arabic exhibits substantial dialectal variation across the Arabic-speaking world, with regional varieties differing in phonological, lexical, and morphosyntactic properties (Versteegh, 2014). This diversity coexists with Modern Standard Arabic (MSA), a supra-regional written norm used in formal domains, while dialects dominate everyday spoken interaction. The resulting diglossic situation (Ferguson, 1959) poses long-standing challenges for descriptive dialectology as well as for Natural Language Processing (NLP), particularly for tasks that require robust modelling of spoken language, such as automatic speech recognition, dialect identification, and cross-dialect transfer (Zampieri et al., 2014; Abdul-Mageed et al., 2020).

Recent years have seen growing interest in Arabic dialect resources and evaluation benchmarks. Datasets such as MADAR (Bouamor et al., 2019) and the NADI shared tasks (Abdul-Mageed et al., 2020) have advanced fine-grained dialect identification, especially for user-generated written text. At the same time, transformer-based Arabic language models such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021) have achieved strong performance across a range of NLP tasks. However, model robustness remains closely tied to the coverage and representativeness of the training data, and dialectal diversity continues to be unevenly captured in many widely used resources (Darwish, 2024). In particular, large-scale spoken corpora that represent multiple dialects and enable systematic comparison across linguistic levels remain limited, restricting progress in speech-oriented modelling and corpus-driven dialect analysis (Ahmed et al., 2022; Al-Shenaifi et al., 2024).

This paper addresses this gap through an empirical analysis of dialectal variation in contemporary Arabic using the SADA corpus (Alotaibi et al., 2024), a large multi-dialectal spoken Arabic resource compiled from diverse media sources and accompanied by speaker and contextual metadata. Rather than focusing on corpus construction, we use SADA as an analytical foundation to quantify and interpret variation across major dialect groups. Our analysis combines quantitative corpus statistics with qualitative linguistic interpretation, including keyword-based comparisons, distributional profiling, and collocation- and trigram-based analyses. These methods allow us to capture variation not only in lexical choice and morphosyntactic tendencies, but also in recurrent phraseological and discourse-pragmatic patterns that characterise dialectal speech.

Using these techniques, we provide evidence that Arabic dialects exhibit systematic differences in lexical selection, morphosyntactic configurations, and frequent multiword expressions, while also sharing a substantial common core. Trigram and collocation patterns further highlight dialect-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 276–286

specific preferences in evaluative language and discourse formulae, which are often underrepresented in resources built primarily from written text. We further compare dialectal distributions with MSA reference data to contextualise divergence and convergence between spoken varieties and the written standard.

The contributions of this work are threefold. First, we present a corpus-based comparative study of Arabic dialects grounded in a large spoken resource, offering empirically testable observations about cross-dialect similarities and differences. Second, we demonstrate the effectiveness of corpus-linguistic methods—particularly those supported by Sketch Engine (Kilgarriff et al., 2014)—for analysing dialectal Arabic at scale. Third, we discuss the implications of dialectal distributional patterns for dialect-aware NLP and speech-related applications, supporting the development of more robust dialect-sensitive models and evaluation practices (Darwish, 2024; Abdul-Mageed et al., 2021).

2 Related Work

Research on Arabic dialectal variation spans descriptive dialectology, sociolinguistics, and Natural Language Processing (NLP), reflecting both the linguistic complexity of Arabic varieties and their increasing importance for language technology. Traditional scholarship documents the historical and geographic stratification of Arabic dialects and highlights the role of diglossia in shaping language use across formal and informal settings (Versteegh, 2014; Ferguson, 1959). In computational research, dialectal diversity has become a central concern due to its impact on model robustness, cross-dialect generalisation, and the performance gap between Modern Standard Arabic (MSA) and spoken varieties.

2.1 Dialectal corpora and resources

Dialectal corpora provide the empirical foundation for Arabic dialectology and dialect-aware NLP. Over the past two decades, substantial progress has been made in developing datasets and benchmarks that support dialect analysis and modelling. For example, MADAR (Bouamor et al., 2019) enables controlled cross-dialect comparison through parallel data spanning 25 city-level dialects, while the Nuanced Arabic Dialect Identification (NADI) shared tasks provide widely used benchmarks for fine-grained dialect identification from social media text (Abdul-Mageed et al., 2020, 2023). For spoken dialects, corpora such as QASR (Mubarak et al., 2021) have supported research using broadcast and conversational speech data.

In addition to broad-coverage resources, specialised corpora have been developed for individual dialects and dialect families. Jarrar et al. (2016) introduced *Curras*, a morphologically annotated corpus for Palestinian Arabic, and Khalifa et al. (2016) developed a large-scale resource for Gulf Arabic. These efforts address gaps in dialect coverage, as many dialects remain underrepresented relative to Egyptian Arabic and MSA-focused resources. Alongside dialect corpora, lexicons and reference corpora such as ArSenL (Badaro et al., 2014), Arabic Gigaword (Agence France-Presse and Linguistic Data Consortium, 2007), and the Quranic Arabic Corpus (Dukes et al., 2011) provide lexical resources and classical or MSA baselines for comparative analysis.

Despite these advances, many widely used benchmarks remain skewed toward written user-generated text, while fewer resources support large-scale comparative analysis across multiple spoken dialects. As a result, speech-prominent phenomena such as pragmatic markers, discourse formulae, and spoken morphosyntactic patterns remain underrepresented in much of the computational literature.

2.2 Dialect identification and computational modelling

A major line of work in Arabic dialect NLP focuses on automatic dialect identification (Zampieri et al., 2014; Abdul-Mageed et al., 2020). Earlier approaches relied on surface lexical cues such as character n-grams and word frequencies, while later work incorporated speech-based representations and learned embeddings. For example, Biadsy et al. (2009) used phonotactic modelling for spoken dialect identification, and Malmasi and Zampieri (2017) applied iVectors and ASR transcripts to distinguish dialects. Other approaches combine acoustic and linguistic features for dialect classification in spoken Arabic (Humayun et al., 2023).

More recently, transformer-based models such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021) have improved performance on dialect identification and other downstream tasks. However, robustness remains highly sensitive to data coverage and dialect distribution, and performance across regional varieties can remain uneven (Darwish, 2024). This challenge is

particularly pronounced for spoken dialects, where transcription variability and speech-specific constructions complicate modelling. In addition, studies on annotation practices show that annotator dialect familiarity can influence data quality and evaluation reliability (Farha and Magdy, 2022).

2.3 Corpus-driven analysis of dialectal variation

Beyond classification, corpus linguistics provides interpretable methods for investigating how dialects differ in systematic ways. Frequency profiling, keyword analysis, collocation methods, and distributional analysis can uncover dialect-specific lexical preferences and recurring phraseological patterns. Work on cross-dialectal Arabic processing emphasises the importance of modelling multi-level variation in NLP pipelines (Harrat et al., 2015), while studies on multi-dialect segmentation further demonstrate the need to account for dialectal variation during preprocessing (Eldesouki et al., 2017). Shared evaluations such as ADI have also provided benchmarks that support comparative study of dialectal signals and model behaviour (Obeid et al., 2019). Nevertheless, corpus-driven analyses that connect lexical and phraseological patterns with interactional and discourse-pragmatic functions remain limited, particularly for large multi-dialect spoken corpora (Ahmed et al., 2022; Al-Shenaifi et al., 2024).

2.4 Positioning of the present study

Within this literature, the present study builds on the SADA corpus (Alotaibi et al., 2024), a large multi-dialect spoken Arabic resource with rich speaker and contextual metadata. In contrast to work that primarily benchmarks dialect classification performance, we use SADA as a foundation for comparative corpus-based analysis across dialect groups. By combining quantitative distributional profiling with collocational and trigram analyses, the study provides interpretable evidence of cross-dialect lexical, morphosyntactic, and discourse-pragmatic variation, supporting both Arabic dialectology and the development of dialect-aware NLP systems.

3 Methodology

The methodology used to analyse Arabic dialectal variation in the SADA corpus consists of four stages: (1) corpus selection and dialectal partitioning, (2) preprocessing and normalisation, (3) corpus-driven feature extraction, and (4) cross-dialect comparison and visualisation. Figure 1 provides an overview of the analytical workflow, while Table 1 reports summary statistics for the dialectal subsets used in the experiments.

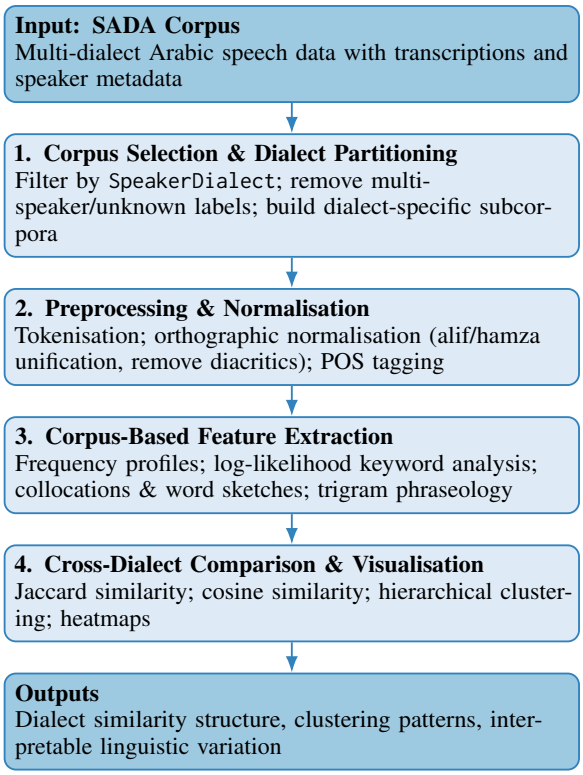

Figure 1: Methodology pipeline for corpus-based analysis of Arabic dialectal variation using the SADA corpus.

3.1 Corpus Selection and Dialect Partitioning

We use the SADA corpus (Alotaibi et al., 2024), a multi-dialectal Arabic speech dataset annotated with transcriptions and speaker metadata. Each speech segment is associated with attributes including *FileName*, *ShowName*, *SpeakerGender*, *SpeakerAge*, and *SpeakerDialect*.

To enable controlled cross-dialect comparison, we filter segments according to the SpeakerDialect label and construct dialect-specific subcorpora for: Najdi, Hijazi, Janubi, Shamali, Khaliji, Egyptian, Levantine, Maghrebi, and MSA. Segments labelled as multi-speaker or with unknown dialect were excluded to avoid confounding effects introduced by mixed speakers or ambiguous dialect assignment.

3.2 Preprocessing and Normalisation

Arabic dialect transcriptions exhibit orthographic variation and inconsistent representations of common characters, which can bias frequency-based and distributional analyses. We therefore apply a

Table 1: Summary statistics of the SADA corpus subsets used for dialectal analysis. The table reports the number of speakers, segments, total duration (in hours), and average segment length (in seconds) per dialect. Segments labelled as multi-speaker recordings were excluded from the comparative analysis.

Dialect	Speakers	Segments	Duration (hrs)	Avg. Segment (s)
Egyptian	6	2,172	2.24	3.72
Hijazi	6	36,170	41.99	4.18
Iraqi	1	2	0.00	1.20
Janubi	6	103	0.09	3.22
Khaliji	6	30,320	31.04	3.69
Levantine	6	966	0.93	3.45
Maghrebi	5	41	0.03	3.03
Modern Standard Arabic	6	4,302	7.96	6.66
Najdi	6	94,611	122.37	4.66
Total	48	168,687	**206.65**	–

normalisation and cleaning pipeline implemented in Python (using pandas, pyarabic, and regex) following established preprocessing practices for Arabic text (Obeid et al., 2023; Ebeid et al., 2023). The pipeline includes:

1. **Tokenisation:** Transcriptions are segmented into word tokens, preserving punctuation.

2. **Orthographic normalisation:** Common variants are mapped to one form (e.g., unifying *alif* and *hamza* variants and removing diacritics).

3. **Morphosyntactic annotation:** Part-of-speech (POS) tagging is performed using Farasa and CamelTools to support later morphosyntactic and collocation-based analyses.

The resulting dialectal subsets are stored as separate text files for analysis in Sketch Engine and Python.

3.3 Corpus-Based Feature Extraction

We perform corpus-driven analysis using Sketch Engine (Kilgarriff et al., 2014) to extract lexical and collocational signals that characterise each dialect. For each subcorpus, we compute:

- **Frequency profiles:** token and type frequency lists as a basis for descriptive statistics and comparison.

- **Keyword analysis:** log-likelihood-based keyness scores are used to identify words that are significantly over-represented in one dialect relative to a reference corpus.

- **Collocations and word sketches:** collocational networks and grammatical relations are extracted to capture dialect-specific phraseological and syntagmatic preferences.

These outputs form the main linguistic feature set used for quantitative comparison and interpretation.

Sketch Engine configuration. All corpus-driven analyses were conducted using Sketch Engine (Kilgarriff et al., 2014). Frequency lists were computed over word forms after preprocessing and normalisation. Keyword analysis employed the log-likelihood statistic with each dialectal subcorpus compared against the remaining combined dialect data as reference. A minimum frequency threshold of 5 occurrences was applied to reduce noise from hapax items. Collocations were extracted using the default Sketch Engine word sketch configuration with grammatical relations based on POS annotation, and a minimum co-occurrence frequency of 5 was required. Trigram extraction was limited to contiguous 3-grams occurring at least 5 times within each subcorpus.

3.4 Cross-Dialect Comparison and Visualisation

To quantify similarity between dialects, we export normalised frequency lists from Sketch Engine and compute cross-dialect similarity measures in Python.

1. **Lexical overlap:** We compute Jaccard similarity between dialect vocabularies to measure shared lexical inventory.

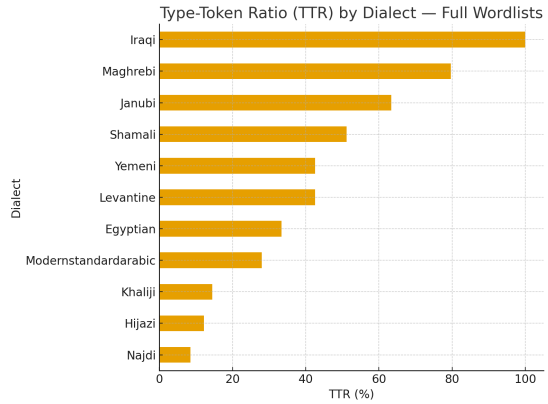

Figure 2: Type–token ratio (TTR) by dialect computed from full wordlists. Larger subcorpora (Najdi, Hijazi) yield lower TTR due to lexical stabilisation with increasing sample size.

2. **Distributional similarity:** We represent each dialect as a frequency vector and compute cosine similarity to capture distributional closeness beyond simple vocabulary overlap.

3. **Clustering and visualisation:** We apply hierarchical clustering to the similarity matrices and visualise dialect groupings using heatmaps implemented with matplotlib and scikit-learn.

This comparison framework enables empirical evaluation of dialect proximity and supports the identification of higher-level dialect clusters (e.g., Gulf vs. Levantine vs. North African groupings).

4 Results

4.1 Corpus Size and Lexical Diversity

Table 2 reports token counts, vocabulary size (types), and type–token ratio (TTR) computed from full wordlists exported from Sketch Engine (Kilgarriff et al., 2014). As expected, TTR decreases with corpus size: large subcorpora such as Najdi and Hijazi show lower TTR values (8–12%), whereas small subcorpora (e.g., Maghrebi, Janubi) exhibit inflated TTR due to limited data and reduced repetition. We therefore interpret TTR primarily as a descriptive indicator of sample size effects rather than as a direct measure of dialectal lexical richness.

Figure 2 visualises this inverse relationship between corpus size and TTR, consistent with well-known corpus-linguistic sampling effects.

4.2 Frequency-Based Lexical Profiles

High-frequency items reveal a substantial shared functional backbone across dialects: common function words such as في, ما, and يا dominate the top ranks in all subcorpora. At the same time, dialect-specific discourse markers and regional forms appear among the most frequent content-bearing items. For example, Hijazi contains markers such as دحين, while Najdi shows frequent items such as زين. These profiles suggest convergence in closed-class vocabulary but divergence in discourse-pragmatic usage, consistent with observations in prior Arabic variation studies (Darwish, 2024).

4.3 Keyword-Based Dialectal Distinctiveness

To quantify lexical distinctiveness, we perform keyword analysis via log-likelihood comparison. Table 3 shows representative top keywords for each dialect. The results highlight salient regional markers such as مره in Hijazi and وش and زين in Najdi, as well as dialect-typical discourse items in Egyptian (e.g., مش, يعني) and Levantine (e.g., شو, بدي, عم). Maghrebi includes region-specific lexemes such as برشا and دبما, while Gulf varieties show frequent temporal/evaluative markers (e.g., الحين, عدل, ماشي). MSA exhibits expected formal items (e.g., ان, الذي). Overall, the strongest distinctiveness signals are concentrated in pragmatic particles, intensifiers, and frequent conversational markers rather than in core function words.

4.4 Trigram Phraseology

Trigram analysis provides a complementary view of dialect identity through recurring multiword units that encode pragmatic style and interactional routines. Table 4 lists representative high-frequency trigrams. Spoken dialects show formulaic conversational patterns (e.g., vocatives, intensification, turn-management phrases), whereas MSA is characterised by formal written sequences (e.g., على الرغم من, في هذا السياق). These differences indicate that dialectal variation is expressed not only through individual words, but also through stable phraseological templates.

Table 2: Lexical diversity statistics across dialectal subcorpora in the SADA dataset (computed from full wordlists).

Dialect	Tokens	Types	TTR (%)
Egyptian	18,053	6,016	33.32
Hijazi	333,193	40,433	12.14
Janubi	726	460	63.36
Khaliji	257,279	37,176	14.45
Levantine	7,497	3,188	42.52
Maghrebi	265	211	79.62
Modern Standard Arabic	52,286	14,605	27.93
Najdi	954,543	80,578	8.44
Shamali	1,638	838	51.16
Yemeni	4,184	1,782	42.59

Table 3: Top 10 dialect-specific keywords identified through log-likelihood analysis (computed on full wordlists). Log-likelihood (LL) values indicate each word's association strength with the dialect relative to other subcorpora.

Dialect	Top 10 Keywords (Word — LL)
Egyptian	مش (491.2), يعني (475.5), قوي (372.8), كده (355.3), ده (301.9), انت (295.1), هنا (271.4), كل (261.7), انا (243.8), على (223.4)
Hijazi	دحين (534.7), كده (501.2), مره (460.8), والله (415.3), اليوم (364.9), شي (337.1), ترى (324.4), بس (315.7), كان (288.2), هذا (269.6)
Najdi	زين (591.0), وش (384.2), يوم (312.1), طيب (289.0), راح (271.5), قلت (251.9), ترى (243.2), بعد (234.5), هناك (212.7), هذا (198.8)
Khaliji	عدل (485.0), الحين (380.0), واجد (305.5), ترى (214.0), بعد (189.0), قال (180.1), زين (172.8), شي (162.3), عندك (157.4), بس (142.7)
Levantine	شو (233.5), بدي (205.0), ابراهيم (194.7), هيك (189.8), بيك (115.6), خيك (107.6), هيدي (107.3), كتير (101.6), بدك (84.2), عم (83.6)
Maghrebi	برشا (322.4), دما (305.8), حاجة (276.9), شي (254.1), قلت (243.5), على (231.6), هذا (217.3), كل (206.2), في (192.9), من (184.4)
Modern Standard Arabic	ان (1180.9), الى (788.2), التي (743.6), يا (742.7), اللي (614.5), بس (612.0), الذي (585.6), ما (519.7), انا (483.2), ذلك (448.3)
Yemeni	ماشي (397.3), قد (355.2), قال (312.7), فين (287.4), خلاص (261.5), انت (240.7), هذا (227.1), انا (216.2), يوم (205.8), من (197.9)
Janubi	شي (254.8), الله (243.6), قال (221.5), يوم (213.2), ترى (198.7), انا (182.1), في (171.3), هذا (160.8), ما (152.4), يا (145.9)
Shamali	والله (289.2), ترى (261.3), يا (247.9), الله (233.4), قال (212.8), بعد (201.7), انا (195.6), في (184.1), هذا (172.5), على (161.8)

4.5 Collocational Patterns

Collocational behaviour further differentiates dialects at the phrase level. Using Sketch Engine word sketches, we observe that the Hijazi intensifier مره systematically collocates with positive evaluative adjectives (e.g., مره حلو), reflecting a productive intensification pattern in this variety. Such collocational regularities provide interpretable signals that complement frequency- and keyword-based distinctiveness.

4.6 Inter-Dialect Similarity and Clustering

We compute cosine similarity (distributional similarity) and Jaccard similarity (vocabulary overlap) from full wordlists to model inter-dialect relationships. Figures 3–5 show that dialect groupings broadly align with geographic proximity: Gulf and Saudi dialects cluster closely, whereas Maghrebi appears comparatively distant from the remaining varieties, consistent with known areal divergence patterns (Darwish, 2024). Finer-grained visualisations (Figures 6–8) further show that Egyptian and Levantine are highly similar relative to other non-Saudi dialects, and that Saudi dialects exhibit strong internal cohesion (cosine > 0.9).

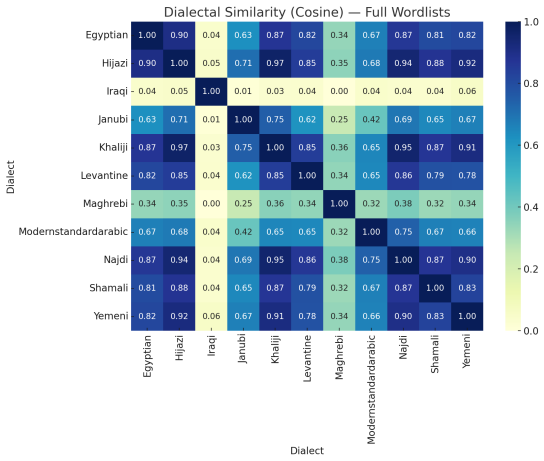

Figure 3: Dialectal similarity (cosine) computed from full wordlists. Gulf and Saudi dialects form a distinct cluster, while Maghrebi diverges sharply.

5 Discussion

Our corpus-based analysis of SADA provides empirical evidence that Arabic dialectal variation is strongly expressed in lexical choice, recurrent phraseology, and discourse-pragmatic routines,

Table 4: Representative high-frequency trigrams across Arabic dialects in the SADA corpus. Trigrams illustrate recurrent phraseological and pragmatic constructions typical of each dialect.

Dialect	Representative High-Frequency Trigrams
Egyptian	يا راجل والله ، مش ممكن والله ، على طول كده ، يعني انت عارف ، كل يوم الصبح ،
	مش عايز اقول ، يا سلام عليك ، خلاص ماشي خلاص ، هو اللي كان ، طب والله العظيم
Hijazi	والله يا شيخ ، بس كده خلاص ، طيب خلاص يلا ، يا رجال والله ،
	ترى والله العظيم ، هو اللي قال ، ما شاء الله ، مره والله حلو ، يا ولد انت ، خلاص انتهى الموضوع
Najdi	الله يرحم والديك ، زين والله العظيم ، ما راح اقول ، وش قلت انت ، يا ولد وش ،
	ترى ما قلت ، بعد شوي يجي ، طيب خلاص خلاص ، انت وش رايك ، الله يوفقك ان شاء
Khaliji	الحين وش تسوي ، عدل والله العظيم ، ما شاء الله ، يا اخي شوف ، انت بعدين روح ،
	ترى يا اخوي ، خلاص عدل تمام ، كل يوم الصبح ، والله الحين قلت ، قالوا لي امس
Levantine	شو يعني انت ، عن جد والله ، بدي روح هلأ ، هيك شي يعني ، ما بعرف بصراحة ،
	والله عن جد ، بدك تروح وين ، انا ما بدي ، كنت عم فكر ، كل يوم الصبح
Yemeni	ماشي الحال خلاص ، قال لي امس ، خلاص يا رجال ، انت قلت لي ، الله يبارك فيك ،
	خلاص تمام خلاص ، ما عاد في ، قد قلت لك ، اليوم قال لي ، ما في داعي
Janubi	الله يرحم والديك ، طيب خلاص انتهى ، ترى قلت لك ، انت وين رايح ، ما قلت لك ،
	يا رجال خلاص ، الله يحفظك ان شاء ، خلاص انتهى الموضوع ، انا قلت لك ، ترى والله العظيم
Shamali	والله يا رجال ، طيب خلاص انتهى ، ما قلت لك ، يا شيخ ترى ، الله يعطيك العافية ،
	ترى والله العظيم ، وش قلت انت ، يا ولد والله ، خلاص تمام خلاص ، الله يرحم والديك
Modern Standard Arabic	قال الله تعالى ، من خلال هذه ، في هذا السياق ، على الرغم من ، الى حد كبير ،
	بناء على ذلك ، في ضوء ما ، في الوقت الذي ، بما ان ذلك ، من جهة اخرى

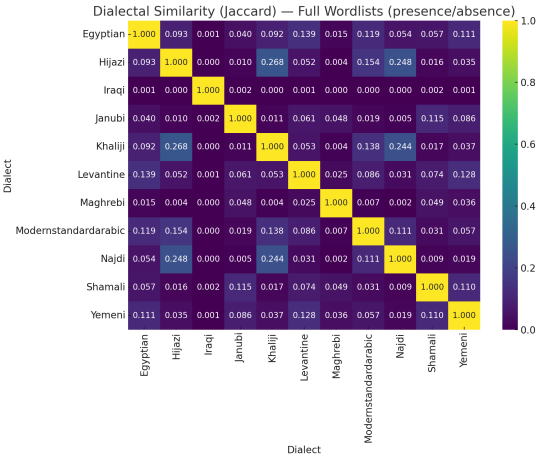

Figure 4: Jaccard similarity (presence/absence) among dialects computed from full wordlists. Similarities broadly align with geographic proximity and sociolinguistic contact.

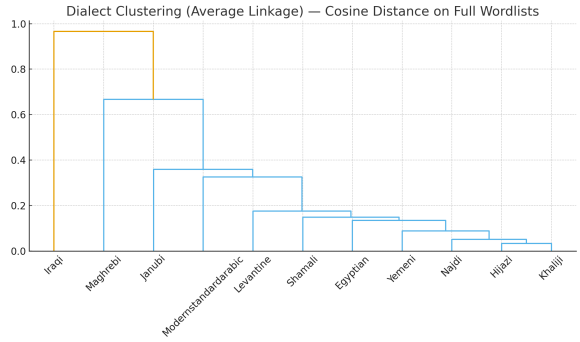

Figure 5: Hierarchical clustering of dialects based on cosine distance (average linkage). Gulf and Saudi varieties cluster closely, reflecting linguistic and geographic continuity.

while sharing a substantial common core across varieties. The similarity structure observed across dialects is broadly consistent with the view of Arabic as a dialect continuum rather than a set of strictly discrete systems (Darwish, 2024). At the same time, the keyword and trigram results show that dialect identity is often carried by high-frequency particles, intensifiers, and formulaic expressions that are central to spoken interaction.

5.1 Dialectal signals beyond core vocabulary

The keyword analysis highlights dialect-salient lexical markers that are widely attested in everyday usage. For example, Egyptian shows high keyness for items such as مش and كده, Hijazi for مره and دحين, and Najdi for زين and وش. Importantly, many of these signals are not content-heavy nouns but rather conversational markers and stance-related items, supporting the interpretation that dialectal distinctiveness is often concentrated in pragmatic and interactional vocabulary.

Trigram patterns complement this view by making explicit the role of formulaic sequences (e.g., بدي روح هلأ, يا راجل والله, والله يا شيخ). These sequences encode common interactional moves such as emphasis, alignment, and turn management, and they provide a linguistically interpretable layer of variation that is not captured by unigram frequencies alone.

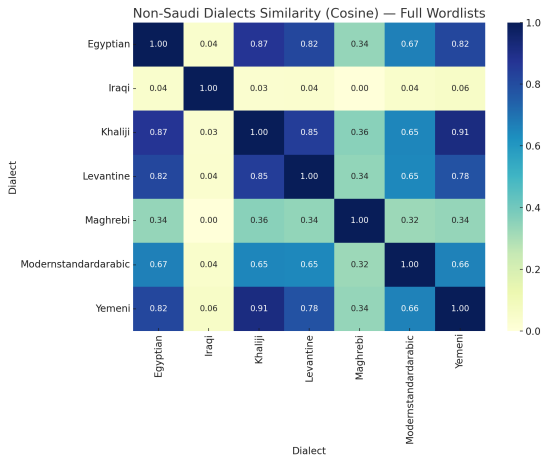

Figure 6: Cosine similarity among non-Saudi dialects using full wordlists. Egyptian and Levantine exhibit high mutual similarity, while Maghrebi diverges.

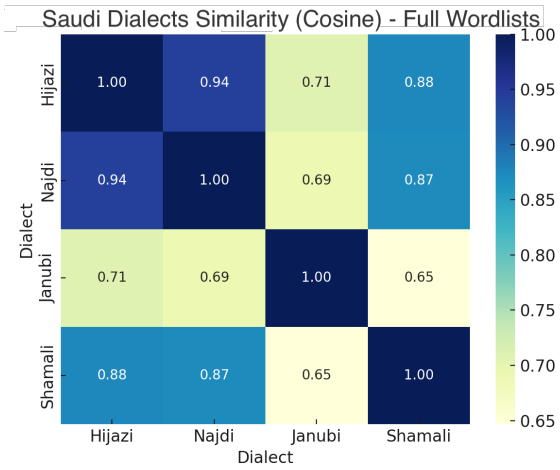

Figure 7: Cosine similarity among Saudi dialects (Hijazi, Najdi, Janubi, Shamali) using full wordlists, showing strong internal cohesion (cosine > 0.9).

5.2 Collocations and dialectal style

Collocational behaviour provides additional evidence that dialects differ not only in what words are frequent, but also in how words combine into recurring constructions. The Hijazi intensifier مره, for instance, exhibits stable collocational associations with evaluative adjectives (e.g., حلو), reflecting a productive intensification pattern. Such collocations provide interpretable signatures of dialectal style and illustrate how spoken dialects often encode stance and expressivity through frequent multiword patterns. Corpus-driven tools such as word sketches support this type of analysis at scale (Kilgarriff et al., 2014).

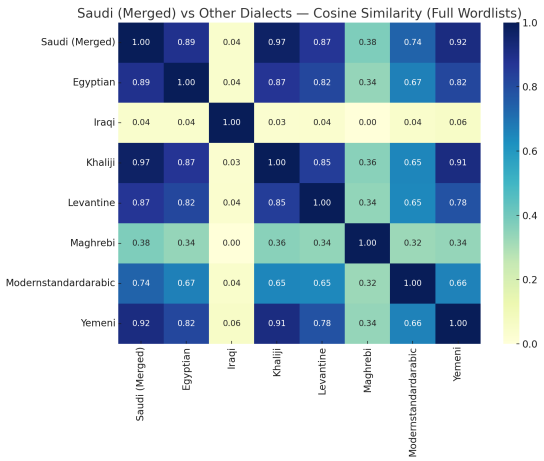

Figure 8: Comparison between merged Saudi dialects and other varieties using full wordlists. Saudi Arabic aligns most closely with Khaliji and MSA, reflecting shared lexical heritage.

5.3 Implications for NLP and machine learning

From an NLP perspective, the results highlight two practical challenges. First, dialectal variation is not confined to low-frequency lexical items: high-frequency discourse markers and formulaic expressions contribute substantially to dialect identity. This implies that dialect-aware modelling should account for frequent multiword units and discourse particles, which are often segmented inconsistently by subword tokenisers and may be underrepresented in MSA-centric training corpora. Second, the similarity and clustering results suggest that dialect groupings align with geographic proximity, which can inform evaluation design (e.g., cross-dialect generalisation tests that reflect realistic transfer settings).

Although transformer-based Arabic models have improved dialect-related classification performance, their effectiveness remains sensitive to the coverage and representativeness of training data (Obeid et al., 2023). The distributional and phraseological patterns observed in SADA therefore motivate incorporating spoken dialectal data into pretraining and evaluation, particularly for speech-related tasks and dialect identification.

5.4 Linguistic interpretation

The contrast between dialectal Arabic and MSA is not only structural but also functional: MSA is primarily used in formal and institutional contexts, whereas dialects dominate conversational interaction. The results presented here suggest that evalua-

tive language, vocatives, and discourse routines are central to dialectal differentiation in speech. These elements encode stance, politeness, and interpersonal alignment, and they help explain why dialect classification can succeed even when dialects share substantial core vocabulary.

5.5 Limitations and Future Work

This study has several limitations. First, dialect representation in the analysed subsets is highly imbalanced, with some varieties (e.g., Najdi, Hijazi) substantially larger than others (e.g., Maghrebi, Janubi). This imbalance has methodological implications for frequency-based and distributional analyses. Larger subcorpora yield more stable estimates of lexical frequencies, collocational behaviour, and keyword distinctiveness, whereas smaller subsets are more susceptible to sampling effects and sparse-data distortions. In particular, limited data may inflate type–token ratios and amplify apparent distinctiveness in keyword or trigram analyses due to reduced repetition. Although similarity measures were computed using normalised frequency vectors to mitigate sensitivity to absolute corpus size, re- sults for underrepresented dialects should be inter- preted with appropriate caution. Future work should incorporate matched-size sampling or controlled re- sampling procedures to systematically evaluate the robustness of dialect similarity patterns under bal- anced conditions.

Second, spoken corpora introduce transcription and orthographic variability that can affect frequency- based comparisons, even after normalisation. While standardisation procedures were applied (e.g., alif and hamza unification, diacritic removal), residual variation may still influence lexical counts and col- locational extraction.

Third, similarity and clustering were computed from full wordlists. Although this approach provides a broad view of distributional proximity, alternative feature selections—such as restricting analyses to shared vocabulary, applying frequency thresholds, or incorporating syntactic or semantic representa- tions—may yield complementary perspectives on dialect relationships.

Future work will extend the analysis to morphosyn- tactic variation using the available POS annotations and will incorporate embedding-based representa- tions to model semantic proximity across dialects. Expanding the use of sociolinguistic metadata (e.g., speaker region, genre, and communicative context) would also enable finer-grained analyses of within- dialect variation and dialect mixing. Overall, the study demonstrates that corpus-driven analysis of spoken dialects can yield interpretable linguistic in- sights while also informing the development of dialect-aware Arabic NLP systems.

6 Conclusion

We presented a corpus-based study of dialectal variation in contemporary Arabic using the SADA corpus. Using frequency profiling, log-likelihood keyword analysis, trigram phraseology, collocational patterns, and similarity-based clustering, we showed that Arabic dialects share a substantial common core while exhibiting systematic differences in lexical choice, multiword expressions, and discourse-pragmatic routines.

Across dialects, the strongest distinctiveness signals were concentrated in high-frequency conversational markers and stance-related items (e.g., مره in Hijazi, مش in Egyptian, زين in Najdi), and in recurrent interactional templates revealed by trigram analysis. Similarity modelling also indicated dialect groupings that broadly align with regional proximity, supporting a data-driven view of Arabic as a continuum with geographically structured variation.

These findings have direct implications for Arabic NLP, particularly for dialect identification and speech-orientated applications. They motivate increased use of spoken multi-dialect resources in model training and evaluation and highlight the importance of capturing frequent multiword units and discourse particles that are central to dialect identity. Future work will extend the analysis to morphosyntactic and prosodic variation and explore embedding-based modelling for cross-dialect similarity under controlled sampling and feature selection.

References

M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi. 2021. Arbert amp; marbert: Deep bidirectional transformers for arabic. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Confer*, pages 7088–7105.

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.

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.

Agence France-Presse and Linguistic Data Consortium. 2007. Arabic gigaword third edition.

A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouani, A. Abd-Alrazaq, and M. Househ. 2022. Freely available arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100049.

Noura Al-Shenaifi and 1 others. 2024. Advancing ai-driven linguistic analysis: Developing and evaluating large-scale arabic dialectal corpora. *Mathematics*, 12(19):3120.

Yousef A. Alotaibi, Sameer Alsharhan, Faisal Alotaibi, Ahmed Alshehri, Mohammad Alghamdi, Mohammed Alqahtani, and Abdulrahman Almuhareb. 2024. Sada: A large-scale multi-dialect arabic speech corpus for linguistic and computational applications. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11716–11720. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*.

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken arabic dialect identification using phonotactic modeling.

H. Bouamor, S. Hassan, and N. Habash. 2019. The madar shared task on arabic fine-grained dialect identification. *Proceedings of the Fourth Arabic Natural Language Processing Workshop*.

Kareem Darwish. 2024. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*.

K. Dukes, E. Atwell, and N. Habash. 2011. Supervised collaboration for syntactic annotation of quranic arabic. *Language Resources and Evaluation*, 47:33–62.

Ahmed Ebeid, Younes Samih, and Kareem Darwish. 2023. A comprehensive study of arabic text normalization techniques for nlp applications. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7)*, pages 45–55, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohamed I. Eldesouki, Younes Samih, Ahmed Abdelalí, Mohammed Attia, Hamdy Mubarak, Kareem Darwish, and Laura Kallmeyer. 2017. Arabic multi-dialect segmentation: Bi-lstm-crf vs. svm.

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of arabic dialect familiarity on data annotation.

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

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, and Kamel Smaıli. 2015. Cross-dialectal arabic processing.

Mohammad Ali Humayun, Hayati Yassin, and Pg Emeroylariffion Abas. 2023. Dialect classification using acoustic and linguistic features in arabic speech. *Iaes International Journal of Artificial Intelligence (Ij-Ai)*.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: An annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf arabic.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: Ten years on. *Lexicography*, 1(1):7–36.

Shervin Malmasi and Marcos Zampieri. 2017. Arabic dialect identification using ivectors and asr transcripts.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Omar Obeid, Go Inoue, and Muhammad Abdul-Mageed. 2023. A benchmark suite for arabic nlp: Dialectal, classical, and modern standard evaluation. *Natural Language Engineering*, 29(5):765–789.

Ossama Obeid, Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2019. ADIDA: Automatic dialect identification for Arabic. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics (Demonstrations), pages 6–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Kees Versteegh. 2014. *The Arabic Language*. Edinburgh University Press.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.