ArabicDialectHub: A Cross-Dialectal Arabic Learning Resource and Platform

Salem Lahlou
Mohamed Bin Zayed University of Artificial Intelligence
UAE
salem.lahlou@mbzuai.ac.ae

Abstract

We present ArabicDialectHub, a cross-dialectal Arabic learning resource comprising 552 phrases across six varieties (Moroccan Darija, Lebanese, Syrian, Emirati, Saudi, and MSA) and an interactive web platform. Phrases were generated using LLMs and validated by five native speakers, stratified by difficulty, and organized thematically. The open-source platform provides translation exploration, adaptive quizzing with algorithmic distractor generation, cloud-synchronized progress tracking, and cultural context. Both the dataset and complete platform source code are released under MIT license. Platform: https://arabic-dialect-hub.netlify.app.

1 Introduction

Arabic, spoken by over 400 million people across diverse regions, exhibits significant dialectal variation that creates substantial barriers to cross-dialectal communication. While Modern Standard Arabic (MSA) serves as a lingua franca in formal contexts, daily communication relies heavily on regional dialects that differ significantly in lexicon, phonology, and syntax. Moroccan Darija, in particular, stands at considerable linguistic distance from both MSA and other Arabic dialects, incorporating substantial Berber, French, and Spanish influences.

Despite the practical importance of cross-dialectal communication, learning resources remain scarce. Most Arabic language learning platforms focus exclusively on MSA (Rani et al., 2023), leaving dialect speakers without adequate tools to learn other varieties. Existing dialectal resources (Bouamor et al., 2018; Zaidan and Callison-Burch, 2011; Meftouh et al., 2015), while valuable for computational linguistics research, are primarily designed for NLP applications rather than language learning. This gap is particularly acute for Darija speakers seeking to communicate with speakers of Levantine or Gulf varieties.

We present ArabicDialectHub, comprising two complementary contributions. First, we introduce a curated collection of 552 phrases across six Arabic varieties: Moroccan Darija, Lebanese, Syrian, Emirati, Saudi, and MSA. The collection was generated using large language models and validated by five native speakers (three Moroccan Arabic and two Lebanese Arabic speakers). Phrases are stratified by difficulty level (beginner, intermediate, advanced) and organized into thematic categories covering communication scenarios from daily greetings to complex idiomatic expressions. Second, we present an open-source interactive learning platform demonstrating the resource's practical utility through multiple learning modalities: a translation hub for phrase exploration, an adaptive quiz system with intelligent distractor generation, progress tracking with cloud synchronization, and cultural context cards highlighting regional sensitivities.

Our contributions address three critical gaps. Unlike large-scale dialectal corpora designed for NLP research, our collection is explicitly structured for pedagogical use with difficulty stratification and contextual usage notes. Unlike existing Arabic learning applications focusing exclusively on MSA, our platform centers on cross-dialectal learning with Darija as default source. Finally, unlike static datasets, we provide a fully functional interactive system validating the resource's utility for real-world language learning. Both the phrase collection and complete platform codebase are released under MIT license.

2 Related Work

2.1 Arabic Dialect Corpora

Significant efforts document Arabic dialectal variation. The MADAR project (Bouamor et al., 2018) provides parallel translations of 2,000 sentences across 25 Arab city dialects along with MSA, English, and French, accompanied by a lexicon cover-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 1–5

ing 1,045 concepts. The Arabic Online Commentary (AOC) dataset (Zaidan and Callison-Burch, 2011) pioneered collection of naturally-occurring dialectal Arabic, harvesting 52 million words from online news comments with 108,000 sentences manually annotated for dialect identification. The Parallel Arabic DIalectal Corpus (PADIC; Meftouh et al., 2015) contributes 2,000 parallel sentences across six dialects including Moroccan, Algerian, Tunisian, Palestinian, and Syrian varieties. Other works include Bouamor et al. (2014).

For Moroccan Darija specifically, recent developments include the Darija Open Dataset (DODa) (Outchakoucht and Es-Samaali, 2024) with approximately 150,000 Darija-English entries, and the Atlaset dataset (AtlasIA, 2024) containing over 155 million tokens for model pretraining. The Darija-Banking corpus (Skiredj et al., 2025) demonstrates LLM-assisted dataset creation viability, using GPT-4 for initial translation with subsequent validation by five native speakers. While these corpora provide valuable resources for computational linguistics, they primarily serve NLP tasks rather than language learning.

2.2 Cross-Dialectal Learning Tools

Despite abundant Arabic learning applications, few address cross-dialectal communication. Mainstream platforms (Duolingo, Rosetta Stone, Pimsleur) focus exclusively on MSA, leaving dialect learners underserved. Recent work on cross-dialectal Arabic translation (Beidas et al., 2025) evaluates large language models on MADAR and QADI datasets, demonstrating growing technological capability but limited deployment in learner-facing applications. Computer-assisted language learning (CALL) research for Arabic (Haniefa and Samsudin, 2025; Bahari et al., 2025) predominantly addresses MSA acquisition, with AI tools struggling on dialects (El Zahraa, 2025).

2.3 LLM-Assisted Dataset Creation

The use of large language models for linguistic dataset creation has gained acceptance in NLP research. Comprehensive surveys document hundreds of LLM-generated datasets across languages and domains (Liu et al., 2024). Critical to this approach is rigorous validation: the CLIcK benchmark for Korean (Kim et al., 2024) employed four native speakers to validate all samples, while LAG-MMLU for Latvian and Giriama (Etori et al., 2025) used native speaker curation ensuring linguistic and cultural relevance despite persistent automatic translation errors. These precedents establish that LLM-assisted generation with native speaker validation represents viable methodology for creating linguistic resources, particularly for lower-resourced varieties.

3 Methodology

3.1 Phrase Collection

3.1.1 Collection Strategy and Design

The phrase collection was designed to serve practical cross-dialectal communication needs while maintaining pedagogical utility. Our selection criteria prioritized frequency (common everyday expressions), utility (practical value for real-world scenarios), and cultural relevance (appropriateness across diverse contexts). The collection spans 18 thematic categories. To accommodate learners at different proficiency levels, phrases were stratified into three difficulty tiers. Additionally, 400 daily conversation sentences provide extended practice with natural dialogues.

3.1.2 LLM-Assisted Generation and Validation

We useed large language models (Claude 3.5 and GPT-4) to generate initial translations, followed by rigorous human validation. Our prompt engineering approach specified detailed dialect characteristics for each target variety, provided contextual information about appropriate register and formality levels, and requested natural conversational translations rather than literal word-for-word renderings. The generation process proceeded iteratively, with initial translations undergoing consistency checking across dialects to identify outliers. Multiple generation attempts were compared for quality assessment, with validators selecting or synthesizing the most natural options.

The 552 phrases underwent review by five native Arabic speakers: three speakers of Moroccan Arabic (Darija) and two speakers of Lebanese Arabic, all fluent in multiple Arabic varieties and MSA. Validators reviewed phrases independently, focusing on naturalness (authentic native speech), accuracy (semantic equivalence), and cultural appropriateness (suitability for indicated contexts).

3.2 Platform Development

To demonstrate the practical utility of our phrase collection, we developed ArabicDialectHub, an

open-source web application providing multiple interaction modalities for cross-dialectal Arabic learning. The platform serves as both proof-of-concept for the resource's pedagogical value and a contribution to the dialectal Arabic learning ecosystem.

3.2.1 System Architecture

The platform employs a modern web architecture optimized for responsive cross-platform access. The frontend utilizes React 18 and TypeScript for type-safe component development with efficient client-side rendering. The backend leverages two specialized cloud services. Clerk handles authentication, supporting email/password credentials with secure session management. This separation of authentication concerns from data management provides security best practices and simplifies user account management. Supabase provides the data layer through a PostgreSQL database with built-in real-time synchronization capabilities and row-level security policies ensuring users can only access their own progress data while maintaining shared read access to the phrase collection.

The database schema comprises three primary tables. The `phrases` table stores all 552 phrases with complete metadata, synchronized from source JSON files during deployment. The `phrase_progress` table tracks individual user mastery status for each phrase, recording correctness counts, mastery flags, and last review timestamps. The `quiz_attempts` table logs all quiz sessions with scores, question counts, configuration parameters (source/target dialects, difficulty filters), and completion timestamps, enabling progress analysis and learning analytics. The platform is deployed on Netlify CDN with continuous integration from GitHub, ensuring automatic updates as the codebase evolves.

3.2.2 Core Learning Features

Translation Hub serves as the primary exploration interface for discovering and learning phrases across Arabic dialects. The hub displays three randomly-selected unmastered phrases simultaneously, encouraging focused attention while maintaining variety. This limited display prevents overwhelming learners while ensuring adequate exposure to new content. Users can expand phrase cards through accordion components to view translations across all six dialects, facilitating cross-dialectal comparison and pattern recognition. Each phrase includes the Darija original in Arabic script with Latin transliteration, literal English translation, and complete translations for Lebanese, Syrian, Emirati, Saudi, and MSA varieties. Each dialectal translation provides Arabic script, romanization for pronunciation guidance, and usage notes explaining contextual appropriateness. The mastery tracking system employs a simple one-click approach where users mark phrases as "mastered" when confident in their understanding. Once marked, phrases are removed from the default rotation but remain accessible through a "show mastered" toggle, allowing review without cluttering the learning interface. All mastery status changes synchronize immediately to Supabase, enabling seamless cross-device learning. The hub includes search functionality for targeted phrase lookup by text content, and filtering options by category and difficulty level. A progress indicator displays real-time mastery percentage, providing immediate feedback on learning advancement.

Quiz System provides active recall practice essential for language retention through two question types with varying cognitive demands. Multiple-choice questions present a phrase in the source dialect with four answer options in the target dialect. The distractor generation algorithm identifies phrases with similar lexical or phonological characteristics from maintained distractor banks. Word-ordering questions present a phrase in the source dialect and challenge users to arrange shuffled words into the correct target dialect sequence. This question type tests syntactic understanding and productive competence rather than mere recognition, requiring learners to actively construct grammatically correct sentences. The shuffling algorithm ensures that word order is sufficiently scrambled to prevent pattern-based guessing. The system provides immediate feedback after each question with green highlights for correct answers and red highlights for incorrect responses, displaying the correct answer.

Progress Tracker visualizes learning metrics across multiple dimensions to support metacognitive awareness and motivation. The dashboard displays overall statistics including total phrases mastered (both absolute count and percentage of the 552-phrase collection) and average quiz performance across all attempts.

Cultural Context Cards acknowledge that effective cross-dialectal communication requires cultural competence alongside linguistic knowledge. Rather than embedding cultural information solely

within individual phrases, dedicated thematic cards provide broader context about regional differences in social norms, communication styles, and cultural sensitivities. Each cultural card presents key points with concrete examples, regional differences across the five dialects, and practical tips for real-world usage. For instance, the greetings card explains that while formal greetings are universally important, the expected response patterns vary (Syrian Arabic often includes more elaborate well-wishing phrases, while Gulf varieties may incorporate more formal religious expressions). This contextualization prevents social mistakes and enhances pragmatic competence.

4 Discussion

4.1 Resource Contribution

The ArabicDialectHub phrase collection addresses a critical gap in cross-dialectal Arabic resources. While existing corpora such as MADAR and PADIC provide valuable parallel data for computational research, our collection offers explicit pedagogical structuring through difficulty stratification, enabling progressive learning, unlike research corpora treating all data uniformly. Second, rich contextual metadata including usage notes, formality indicators, and cultural sensitivities directly supports learner needs beyond mere translation equivalence.

4.2 Platform Validation

The platform demonstrates that our phrase collection effectively supports interactive language learning. The translation hub validates that the resource structure facilitates browsing and discovery. The quiz system's successful distractor generation confirms sufficient lexical diversity for automated assessment without external word lists. Progress tracking proves feasible through the structured data model. The open-source release establishes a blueprint for dialectal learning applications beyond Arabic.

5 Conclusion

We have presented ArabicDialectHub, a comprehensive resource for cross-dialectal Arabic learning comprising a curated phrase collection and an interactive learning platform. The collection of 552 phrases across six Arabic varieties addresses the critical gap in pedagogically-oriented dialectal resources, particularly for Moroccan Darija speakers. Our methodology combining LLM-assisted generation with native speaker validation demonstrates a viable approach for efficient resource creation in lower-resourced language varieties. The open-source platform provides multiple learning modalities (browsing, quizzing, progress tracking, and cultural context) validating the resource's practical utility.

While acknowledging significant limitations in validation scope, empirical evaluation, and scale, we position this work as an enabling contribution. The complete dataset and platform code are publicly available under the MIT license at https://github.com/saleml/arabic-dialect-hub, with documentation supporting extension and adaptation. We welcome contributions from the research community and language learning practitioners to refine translations, expand dialect coverage, integrate audio resources, and conduct rigorous pedagogical evaluation. By lowering barriers to cross-dialectal Arabic learning and establishing open infrastructure for collaborative development, we aim to support both learner communities and future research on dialectal language education.

Limitations

Validation Coverage. Native speaker validation was limited to Moroccan Darija (three validators) and Lebanese Arabic (two validators). Syrian, Emirati, and Saudi translations lack native speaker verification. Inter-annotator agreement metrics were not computed.

Dataset Scale and Scope. At 552 phrases, the collection remains modest compared to research corpora like MADAR (2,000 sentences). Several domains including technical, medical, and professional vocabulary are absent, constraining utility for specialized communication needs.

Methodology Documentation. Difficulty levels were assessed by Claude 3.5 without formal operationalization criteria. The choice of general-purpose LLMs over Arabic-focused models (e.g., Jais) was driven by availability rather than systematic comparison.

Evaluation. No user studies or learning outcome assessments were conducted. Platform effectiveness for language acquisition remains unvalidated.

Modality. The absence of audio recordings limits pronunciation learning and productive skill development.

Ethics Statement

LLM Use. Claude 3.5 and GPT-4 generated initial translations, with all content undergoing mandatory native speaker validation. LLMs served as productivity tools rather than authoritative sources. Potential biases include systematic preference for formal registers in MSA-trained models.

Data Privacy. Platform collects minimal user data: authentication credentials (Clerk-managed), learning progress (Supabase with row-level security), quiz histories. No PII beyond email retained.

References

AtlasIA. 2024. AL Atlas: Moroccan Darija pretraining. Accessed: 2025-10-07.

Akbar Bahari, Feifei Han, and Artur Strzelecki. 2025. Integrating call and aiall for an interactive pedagogical model of language learning. *Education and Information Technologies*, pages 1–29.

Ayah Beidas, Fatme Ghaddar, Kousar Mohi, Imtiaz Ahmad, and Sa'Ed Abed. 2025. Cross-dialectal arabic translation: comparative analysis on large language models. *Frontiers in Artificial Intelligence*, 8:1661789.

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.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and 1 others. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Fatima El Zahraa. 2025. Leveraging artificial intelligence and digital technologies to enhance sociolinguistic competence and arabic language skills. In *Proceeding of the International Conference on Religious Education and Cross-Cultural Understanding*, volume 1, pages 33–49.

Naome A Etori, Arturs Kanepajs, Kevin Lu, and Randu Karisa. 2025. Lag-mmlu: Benchmarking frontier llm understanding in latvian and giriama.

Rifda Haniefa and Mohamad Samsudin. 2025. The effectiveness of web-based computer assisted language learning in improving arabic speaking skills: Efektivitas computer assisted language learning berbasis web dalam menigkatkan keterampilan berbicara bahasa arab. *Ta'limi Journal of Arabic Education and Arabic Studies*, 4(1):59–72.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for large language models: A comprehensive survey.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of darija open dataset: Introducing version 2.

Samsuar A Rani, Zikriati Zikriati, Aan Muhammady, Syukran Syukran, and Banta Ali. 2023. Arabic language learning based on technology (opportunities and challenges in the digital era). *International Journal of Education, Language, and Social Science*, 1(1):1–11.

Abderrahman Skiredj, Ferdaous Azhari, Ismail Berrada, and Saad Ezzini. 2025. Darijabanking: A new resource for overcoming language barriers in banking intent detection for moroccan arabic speakers. *Natural Language Processing*, 31(5):1234–1264.

Omar 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.