# Tarab: A Multi-Dialect Corpus of Arabic Lyrics and Poetry

**Mo El-Haj**

VinUniversity, Vietnam

Lancaster University, UK

`elhaj.m@vinuni.edu.vn`

`m.el-haj@lancaster.ac.uk`

## Abstract

We introduce the Tarab Corpus, a large-scale cultural and linguistic resource that brings together Arabic song lyrics and poetry within a unified analytical framework. The corpus comprises 2.56 million verses and more than 13.5 million tokens, making it, to our knowledge, the largest open Arabic corpus of creative text spanning both classical and contemporary production. Tarab is broadly balanced between songs and poems and covers Classical Arabic, Modern Standard Arabic (MSA), and six major regional varieties: Egyptian, Gulf, Levantine, Iraqi, Sudanese, and Maghrebi Arabic. The artists and poets represented in the corpus are associated with 28 modern nation states and multiple historical eras, covering over fourteen centuries of Arabic creative expression from the Pre-Islamic period to the twenty-first century. Each verse is accompanied by structured metadata describing linguistic variety, geographic origin, and historical or cultural context, enabling comparative linguistic, stylistic, and diachronic analysis across genres and time. We describe the data collection, normalisation, and validation pipeline and present baseline analyses for variety identification and genre differentiation. The dataset is publicly available on HuggingFace at `https://huggingface.co/datasets/drelhaj/Tarab`.

## 1 Introduction

Arabic is characterised by rich linguistic variation across geography, social context, and historical period. Modern Arabic exists in a continuum between Modern Standard Arabic (MSA) and diverse regional dialects, each with distinct phonological, morphological and lexical properties (Habash, 2010). Dialectal Arabic has received increasing attention in recent years due to its prevalence in real-world communication and the limitations of resources focused only on MSA (Zaidan and Callison-Burch, 2014; El-Haj et al., 2018; Bouamor et al., 2018). However, most existing Arabic corpora are drawn from news, Wikipedia, or social media, leaving creative forms of language such as song lyrics and poetry significantly underrepresented (Attia et al., 2008; Obeid et al., 2020; El-Haj and Ezzini, 2024).

Song lyrics and poetry are valuable for Arabic NLP because they encode features that are often absent from standard corpora, including rhyme, metre, emotional expression, repetition, discourse parallelism, and genre-specific conventions. These genres frequently include dialectal intensity, morphological variation, and non-standard orthography (Darwish, 2013; Habash, 2010), as well as code-switching between Arabic varieties and other languages (Habash et al., 2014; El-Haj and Ezzini, 2024). Poetry also captures Classical Arabic forms across historical eras, offering opportunities for diachronic linguistic analysis (Al-Shaibani et al., 2020; Qarah, 2024). Despite this linguistic richness, there is currently no large-scale, publicly available corpus that unifies both Arabic song lyrics and poetry in a way that supports comparative analysis across dialects, genres, and historical periods.

This paper introduces the Tarab Corpus, a large-scale resource of Arabic creative language encompassing both song lyrics and poetry across modern and historical contexts[1]. Tarab, often translated as musical ecstasy or aesthetic rapture, refers to a culturally grounded affective state of deep emotional engagement experienced in Arabic musical and poetic tradi-

---

[1] `https://huggingface.co/datasets/drelhaj/Tarab`

tions. The corpus comprises 2,557,311 verses and 13,509,336 tokens, with each verse annotated for linguistic variety, geographic origin, and historical or cultural context. Tarab spans texts from contemporary popular music and modern poetry to classical literary traditions associated with major historical eras, capturing Arabic language use across time, region, and genre. In contrast to existing resources that are typically restricted to a single variety or domain (Zaidan and Callison-Burch, 2014; Bouamor et al., 2018), Tarab enables cultural, computational, and sociolinguistic research at a scale and level of diversity not previously available.

## 2 Related Work

Arabic language resources have expanded in recent years, yet most available corpora focus on news and encyclopaedic text (El-Haj and Koulali, 2013; Antoun et al., 2020). Major efforts such as the Arabic Gigaword Corpus (Parker et al., 2011) and the OSIAN web corpus (Zeroual et al., 2019) support large-scale modelling of Modern Standard Arabic (MSA) but do not address dialectal or creative linguistic forms. With the rise of interest in Arabic dialect processing, several dialectal corpora have been introduced, including the Arabic Online Commentary dataset (Zaidan and Callison-Burch, 2011), the MADAR corpus of parallel sentences across Arabic cities (Bouamor et al., 2018), and country-level social media corpora (Alhazmi et al., 2024; Abdelali et al., 2020). These resources enabled progress in dialect identification, but they are limited to prose and do not represent verse or musical language.

Work on Arabic poetry and cultural text remains relatively scarce in NLP. Projects such as OpenITI (Padillo-Saoud, 2019) and Shamela (Belinkov et al., 2016) have made important progress in digitising classical Arabic texts, and the AlKhalil morphological analyser for Classical Arabic (Boudlal et al., 2010; Boudchiche et al., 2017) enables heritage text analysis. However, these collections focus primarily on prose rather than poetry or song. Arabic poetry has been studied computationally in the context of metre classification (Al-Shaibani et al., 2020; Mutawa and Alrumaih, 2025), but

available datasets are small in scale and constrained to classical forms. There remains a gap in large unified poetic corpora that also include modern verse and dialectal variation.

Song lyrics represent another creative domain that reflects informal language and dialectal richness, but they are significantly underrepresented in Arabic NLP. Lyrics exhibit features such as rhyme, repetition and colloquial morphology, making them useful for studying linguistic variation and stylistic modelling. El-Haj (2020) introduced the Habibi Lyrics Corpus , one of the first Arabic lyrics resources covering multiple dialects. That work demonstrated the value of lyrics for dialect identification, but it was limited to musical content and did not include poetry or historical linguistic dimensions.

The Tarab Corpus builds on this line of research by extending the scope of creative Arabic resources beyond lyrics to also include poetry. Unlike previous datasets, Tarab integrates both modern and classical text, linking verse-level entries to dialect, origin and historical metadata. This makes it possible to study variation across genre, geography and historical period within a single framework. To our knowledge, this is the first Arabic resource to unify lyrics and poetry at scale for linguistic, cultural and computational analysis.

## 3 Corpus Creation and Design

The Tarab Corpus is a large-scale resource of Arabic creative expression that brings together song lyrics and poetry within a single, unified framework. Rather than treating these genres as separate cultural artefacts, Tarab adopts the verse as its basic unit of analysis, enabling systematic comparison across genre, linguistic variety, geography, and historical period. This design supports analyses that span performance, literature, and orality, which are difficult to conduct using existing Arabic resources that focus primarily on prose or single varieties.

Tarab captures both contemporary and heritage forms of Arabic creativity. It combines a broad spectrum of song lyrics drawn from popular, folk, and religious repertoires with a substantial body of Arabic poetry ranging from early literary traditions to modern poetic practice. In total, the corpus comprises 2,557,311

verses and more than 13.5 million tokens, representing 89,166 distinct works produced by 2,598 unique creators (2,060 singers and 538 poets) associated with 28 modern countries and major historical eras, from the Pre-Islamic period through successive Islamic dynasties to the present. Linguistic coverage spans Classical Arabic, Modern Standard Arabic (MSA), and six major regional dialect groups, supporting research that connects Arabic literary heritage with contemporary popular culture.

Tarab is constructed from three main streams. First, the poetry component builds on an openly available Arabic poetry collection released on Kaggle[2]. Second, the lyrics component includes material from the Habibi corpus (El-Haj, 2020). Third, we extend coverage by crawling additional publicly accessible web pages containing lyric text. Crawling was restricted to sites that permit automated access, operationalised by checking that the site's `robots.txt` does not disallow retrieval of the relevant paths. The final dataset is represented uniformly at verse level, with all sources normalised into the same schema described in Section 3.2.

## 3.1 Creative scope

Tarab draws from two primary creative domains: song lyrics and poetry. The lyrics component spans a wide range of stylistic and cultural contexts rather than fixed, explicitly annotated genres. These include mainstream popular songs, religious (*dīnī*) material, hip-hop and rap, and songs associated with particular musical styles or performance traditions such as Khaleeji and Maghrebi. These stylistic categories are not treated as mutually exclusive labels tied to artist nationality or dialect. For instance, a song may be performed by an artist from Tunisia, contain Maghrebi dialectal features, and yet follow a Khaleeji musical style. Such distinctions are preserved through separate metadata fields and auxiliary resources rather than collapsed into a single genre label.

The poetry component includes both contemporary poetry and heritage poetry. Contemporary poems are associated with modern national origins, such as Iraq, the United Arab Emirates, or Palestine, while heritage poetry is

linked to major historical periods including the Abbasid, Ayyubid, Andalusian, Mamluk, and Ottoman eras. This dual representation enables the study of poetic language across both modern sociocultural contexts and long-term historical trajectories. Together, the two domains provide a continuous view of Arabic creative language across performance traditions, registers, and time, while allowing dialect, style, and origin to be analysed independently.

## 3.2 Verse-level representation and schema

All content in Tarab is represented using a unified verse-level schema. Each verse occupies a single row and is linked to its parent work through stable identifiers, allowing both fine-grained linguistic analysis and reconstruction of full songs or poems when needed. The schema includes the following fields: `art_id`, `artist_id`, `artist_name`, `art_title`, `writer`, `composer`, `verse_order`, `verse_lyrics`, `origin` (modern country or historical era), `dialect`, and `type` (song or poem). This representation supports longitudinal analysis, cross-genre comparison, and reproducible experimentation across linguistic varieties and historical periods.

## 3.3 Pre-processing

All text in Tarab is stored in UTF-8 and undergoes minimal pre-processing in order to preserve dialectal, orthographic, and stylistic variation. Orthographic features that carry linguistic or regional signal, such as Egyptian *alef maqṣūra* usage, Gulf vowel elongation, and Maghrebi conventions, are intentionally retained. Verse segmentation follows the line structure of the source material, and the `verse_order` field preserves intra-song and intra-poem sequencing. No stemming, lemmatisation, or stopword removal is applied, avoiding the loss of information relevant to linguistic, stylistic, and cultural analysis.

To ensure internal consistency and prevent duplication, works are identified and validated using a composite key defined over (`art_id`, `artist_id`, `verse_order`). This allows repeated verses, alternative textual witnesses, and variant performances to be handled systematically while preserving a clear notion of what constitutes a distinct creative work.

---

## 3.4 Corpus composition and growth

Table 1 summarises the composition of the Tarab corpus by genre. While poetry accounts for a larger share of works and verses, the two genres differ in average verse length, reflecting stylistic differences between poetic and musical forms.

Tarab represents a substantial expansion over earlier Arabic lyrics resources. Compared to the Habibi corpus (El-Haj, 2020), which contains 527,896 lyric verses, Tarab increases the total number of verses by a factor of 4.8, incorporating an additional 1,387,283 verses of poetry alongside 642,221 further lyric verses. This expansion broadens the scope from purely modern musical texts to a unified collection spanning contemporary songwriting and classical Arabic poetics. The `corpus_version` field indicates whether a song was originally present in the Habibi corpus, supporting controlled analyses of diachronic and genre variation. Habibi corpus did not include poetry. Figure 1 illustrates the relative contribution of each corpus version.

| Subset | Works | Verses | Tokens | Avg tokens/verse |
|--------|-------|--------|--------|------------------|
| Songs | 34,239 | 1,170,028 | 6,989,019 | 4.9 |
| Poems | 54,927 | 1,387,283 | 6,520,317 | 5.6 |
| **Total** | **89,166** | **2,557,311** | **13,509,336** | 5.3 |

Table 1: Composition of the Tarab corpus by genre, showing the number of works, verses, tokens, and average verse length.
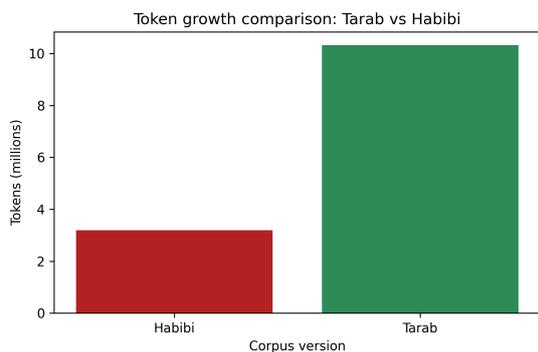


Figure 1: Corpus growth in scale compared to earlier Arabic lyrics resources.

# 4 Linguistic, Geographic, and Structural Coverage

This section describes the linguistic, geographic, and structural properties of the Tarab corpus, followed by a detailed analysis of its lexical and stylistic characteristics. Together, these perspectives provide a comprehensive account of how Arabic creative language is distributed, structured, and realised across dialects, genres, and historical contexts.

## 4.1 Linguistic and dialectal coverage

At the linguistic level, Tarab spans Classical Arabic, Modern Standard Arabic (MSA), and six major regional dialect groups. Table 2 summarises the distribution of verses by dialect, together with vocabulary size and average verse length. Classical Arabic and MSA together account for a substantial proportion of the corpus, reflecting the prominence of poetry and formal literary production. In contrast, song lyrics contribute extensive coverage of spoken regional varieties, including Egyptian, Gulf, Levantine, Iraqi, Sudanese, and Maghrebi Arabic, ensuring that contemporary vernacular usage is well represented.

| Dialect | Verses | Vocab size | Avg tokens/verse | % of corpus |
|---------|--------|-----------|------------------|-------------|
| Classical | 937,473 | 1,044,325 | 4.7 | 36.7 |
| MSA | 449,810 | 577,073 | 4.6 | 17.6 |
| Egyptian | 308,714 | 120,507 | 6.3 | 12.1 |
| Gulf | 308,249 | 133,599 | 6.1 | 12.1 |
| Levantine | 250,276 | 119,455 | 5.9 | 9.8 |
| Iraqi | 156,153 | 73,531 | 5.5 | 6.1 |
| Sudanese | 89,226 | 58,092 | 5.7 | 3.5 |
| Maghrebi | 57,410 | 33,762 | 6.0 | 2.2 |

Table 2: Vocabulary size and average verse length by dialect in the Tarab corpus.

## 4.2 Geographic and historical provenance

Tarab incorporates material associated with both modern nation states and major historical eras, spanning over fourteen centuries of Arabic creative text, from pre-610 CE poetry to contemporary songs and modern poetic production in the twenty-first century. Figure 2 shows the most prominent origins by verse count. Modern countries such as Egypt, Lebanon, and Saudi Arabia contribute a large share of song lyrics, while historical periods including the Abbasid, Andalusian, and Mamluk eras account for a substantial proportion of the poetic material. This explicit separation between geographic origin and historical era enables analysis across time and space without conflating linguistic variety with chronology.

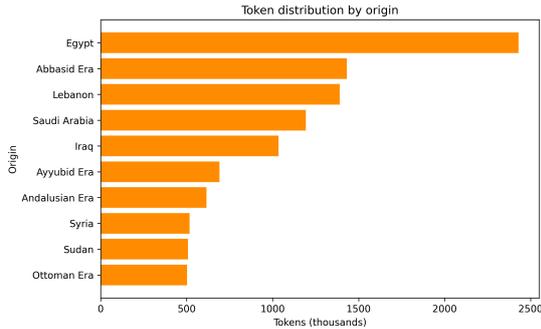Table 3 presents the full distribution of

Figure 2: Top origins by verse count, including modern countries and historical eras.
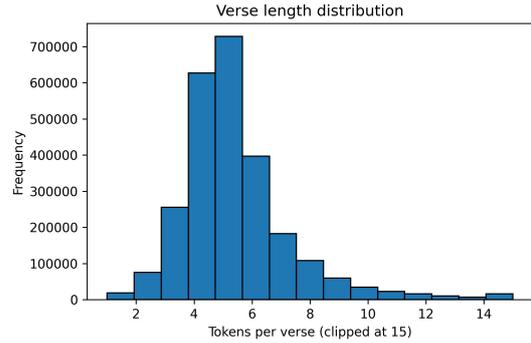


Figure 3: Distribution of verse lengths across the Tarab corpus.

works, tokens, and verses across modern countries and historical eras.

| Origin | Works | Tokens | Verses |
|---|---|---|---|
| Egypt | 11,182 | 2,429,198 | 414,914 |
| Abbasid Era | 13,456 | 1,431,613 | 303,378 |
| Lebanon | 7,390 | 1,390,369 | 253,143 |
| Saudi Arabia | 6,575 | 1,193,549 | 197,384 |
| Iraq | 4,913 | 1,034,427 | 195,165 |
| Ayyubid Era | 5,018 | 690,972 | 143,768 |
| Andalusian Era | 4,410 | 616,022 | 130,040 |
| Ottoman Era | 3,937 | 502,892 | 108,743 |
| Mamluk Era | 6,095 | 490,866 | 102,999 |
| Syria | 2,820 | 517,833 | 99,693 |
| Sudan | 2,683 | 507,783 | 89,829 |
| Kuwait | 1,962 | 361,052 | 61,867 |
| Palestine | 1,429 | 271,712 | 56,448 |
| United Arab Emirates | 1,719 | 310,004 | 54,462 |
| Islamic Era | 2,351 | 264,482 | 54,081 |
| Morocco | 1,259 | 235,739 | 41,298 |
| Era of the Mukhadramun | 2,167 | 192,953 | 40,692 |
| Pre-Islamic Era | 1,989 | 175,622 | 36,826 |
| Tunisia | 1,072 | 168,709 | 31,671 |
| Yemen | 1,360 | 153,797 | 30,535 |
| Algeria | 807 | 129,197 | 25,157 |
| Umayyad Era | 2,360 | 124,200 | 24,817 |
| Jordan | 775 | 125,656 | 23,574 |
| Oman | 872 | 95,100 | 19,872 |
| Bahrain | 207 | 35,515 | 5,863 |
| Qatar | 199 | 33,696 | 5,723 |
| Libya | 133 | 18,292 | 3,775 |
| Mauritania | 27 | 8,086 | 1,594 |
| **Total** | **89,166** | **13,509,336** | **2,557,311** |

Table 3: Distribution of works, tokens, and verses across modern countries and historical eras.

### 4.3 Structural properties of verses

At a structural level, verses in Tarab are typically short. Figure 3 shows the distribution of tokens per verse, with most verses falling between three and eight tokens.

### 4.4 Dialectal lexical variation

Given this distributional profile, Tarab exhibits linguistic behaviour that differs markedly from newswire and social media corpora commonly used in Arabic NLP. Lexical choice and tokenisation patterns are shaped by creative constraints, including metre, repetition, and per-

formance, rather than sentence-based prose structure.

Dialectal variation is particularly visible in vocabulary composition. Classical Arabic and MSA display the largest vocabularies, consistent with the lexical richness and stylistic range of poetic language. Regional dialects, while smaller in vocabulary size, exhibit strong lexical distinctiveness and longer average verse lengths, especially in song lyrics.

Beyond aggregate statistics, Tarab exhibits clear and systematic dialectal differentiation that reflects regionally grounded usage across the corpus. This diversity is evident in the high-frequency lexical items summarised in Table 4, which highlight recurrent dialect-specific forms rather than shared pan-Arabic vocabulary. Across regional varieties, these items include characteristic discourse particles, address forms, and affective expressions that are widely attested in spoken interaction and creative language.

For instance, Maghrebi varieties show frequent use of forms such as علاش (why), بغيت (I want), and مازال (still), which are strongly associated with Maghrebi Arabic. Similarly, Gulf Arabic is characterised by vocative and affective expressions such as وينك (where are you) and يا بعدي (my beloved), while Egyptian and Levantine varieties exhibit colloquial particles and pronominal forms typical of everyday speech. Together, the patterns illustrated in Table 4 demonstrate that Tarab captures robust dialectal influence across regions, strengthening the corpus's linguistic diversity and its suitability for research on dialect modelling and regional stylistics.

41

| Dialect | Common lexical items | Gloss (English) |
|---|---|---|
| Classical | قال، يا، إن، على، ما | said, O/oh, indeed, on/upon, what |
| MSA | الذي، هذه، كان، إلى، عن | which/that, this, was, to, about |
| Egyptian | ليه، بس، قلبي، ده، إنت | why, just, my heart, this, you |
| Gulf | ويك، يا بعدي، أحبك، الغلا | where are you, my beloved, I love you, dear |
| Levantine | شو، حبيبي، هيك، لسه، قلبي | what, my love, like this, still, my heart |
| Iraqi | شلون، آني، روحي، إنت، وين | how, I, my soul, you, where |
| Sudanese | خلاص، مالك، حبابك، وين، سلام | enough/ok, what's wrong, welcome, where, peace |
| Maghrebi | بغيت، مازال، علاش، حبي، قلبي | I want, still, why, my love, my heart |

Table 4: Examples of frequent lexical items by dialect with English glosses.

## 4.5 Code-switching and multilingual influence

Tarab contains natural but unevenly distributed instances of code-switching, overwhelmingly concentrated in song lyrics. Code-switching occurs in approximately 0.6% of song verses and is virtually absent in poetry. At the artwork level, around 2.3% of songs contain at least one instance of code-switching, compared to fewer than 0.1% of poems. Latin-script tokens account for about 0.44% of all song tokens and are negligible in poetry.

The code-switched material consists primarily of French and English lexical items, particularly in Maghrebi and Lebanese lyrics, including *mon amour*, *baby*, *merci*, and *fiesta*. These patterns align with contemporary sociolinguistic practice in popular music and highlight Tarab's value for studying multilingualism and language contact in Arabic creative contexts.

## 4.6 Word-level lexical structure

Beyond aggregate statistics, Tarab enables fine-grained analysis of how lexical items associated with different varieties and genres are organised in distributional space. To explore this, we conduct a word-level analysis using FastText embeddings trained on the Tarab corpus. Focusing on word types rather than verses or documents allows us to examine lexical relationships directly, without conditioning on higher-level structural or stylistic units. This is particularly relevant for Arabic, where variation across dialects and registers is often realised at the lexical and morphological level.

FastText (Bojanowski et al., 2017) is well suited to this setting, as its subword modelling captures morphological variation and orthographic regularities characteristic of both standard and non-standard varieties of Arabic. We retain the full vocabulary when training and analysing the embeddings, allowing frequent and infrequent items alike to contribute to the structure of the space. The resulting word embeddings are projected into two dimensions using t-SNE to support qualitative inspection of how lexical items associated with different varieties and genres are distributed within a shared embedding space.
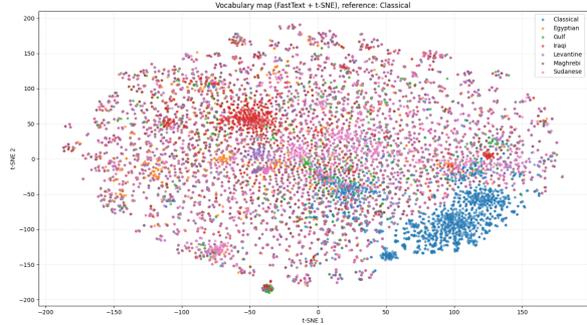


Figure 4: Word-level vocabulary map with Classical Arabic as the reference variety.

Figure 4 visualises the lexical space with Classical Arabic as the reference variety. Classical Arabic forms a compact and largely isolated region, with minimal overlap with dialectal vocabularies. This pattern is consistent with the specialised and genre-bound use of Classical Arabic in Tarab, where its vocabulary tends to occur in constrained poetic and rhetorical contexts that are rarely shared with colloquial varieties.

In contrast, Figure 5 shows Modern Standard Arabic (MSA) occupying a denser and more permeable core of the lexical space. While MSA vocabulary remains internally cohesive, dialectal word forms are distributed around and partially interleaved with it, suggesting substantial lexical sharing and contextual proximity. This organisation aligns with the role of MSA in Tarab as a central written and semi-formal register that co-exists with regional dialects, particularly in song lyrics.

It is important to note that this analysis does not explicitly distinguish between poetic texts and song lyrics. While Classical Arabic in Tarab is predominantly realised in poetry, and MSA material, though often poetic in form, is frequently performed in songs, these genre differences are not encoded in the embedding space and are therefore conflated in the visualisation.

Figure 6 contrasts poems and song lyrics at the word level. The visualisation shows a clear
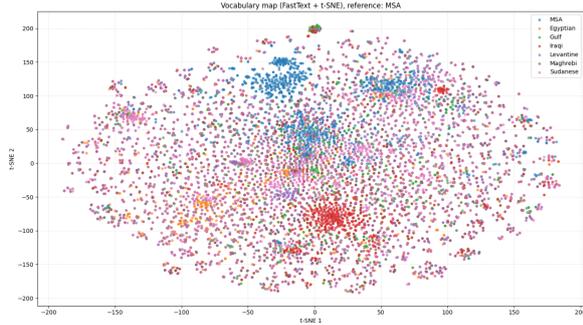
Figure 5: Word-level vocabulary map with MSA as the reference variety.

separation between poetic and lyrical vocabularies once shared high-frequency items are removed. Poetic vocabulary forms a compact and internally cohesive region, consistent with conventionalised lexical choices associated with literary poetry. In contrast, song vocabulary occupies a broader and more fragmented region of the space, suggesting greater lexical diversity and the coexistence of multiple expressive strategies shaped by performance, repetition, and colloquial usage.

Taken together, these visualisations point to a layered lexical structure in Tarab: a highly distinct Classical Arabic stratum, a central and connective MSA layer, and regional dialects and song-specific vocabularies that combine shared lexical material with clusters of strongly distinctive items. This structure highlights the potential of Tarab as a resource for studying lexical variation, register interaction, and dialect-aware representation learning in Arabic NLP.



Figure 6: Word-level vocabulary map contrasting poems and song lyrics.

## 5   Artist and Poet Coverage

The Tarab corpus includes 34,239 unique song titles and 54,927 unique poem titles, reflecting the cultural diversity of Arabic musical and poetic heritage across both modern and historical contexts. In contrast to the original Habibi corpus (El-Haj, 2020), which ranked artists using raw verse frequency alone, Tarab adopts a more balanced ranking approach that accounts for multiple dimensions of contribution. This reduces bias toward prolific artists with short or repetitive works, as well as poets with unusually long or formulaic compositions.

Specifically, we compute a composite contribution score that equally weights three factors: productivity (number of songs or poems), textual volume (total word count), and dataset presence (total number of verses). Together, these measures capture both the breadth and depth of an artist's or poet's contribution to the corpus. The score for each artist or poet is computed as:

$$\text{score} = \tfrac{1}{3} \left( \frac{\text{words}}{\max(\text{words})} + \frac{\text{verses}}{\max(\text{verses})} + \frac{\text{works}}{\max(\text{works})} \right)$$

where *works* refers to songs for lyric artists and poems for poets. Tables 5 and 6 list the most prominent contributors according to this balanced score. The rankings reveal a mixture of modern music figures, such as *Fayrūz* and *Muḥammad ᶜAbdu*, alongside canonical poets from the Abbasid and Ottoman periods, including *al-Sharīf al-Raḍī* and *Abū al-ᶜAlāʾ al-Maᶜarrī*. This distribution highlights the cultural depth of the Tarab corpus and its suitability for research in diachronic stylistics, authorship studies, and cultural analytics.

| Artist | Songs | Words | Verses |
|---|---|---|---|
| فيروز (Fayrūz) | 681 | 124,268 | 21,920 |
| محمد عبده (Muḥammad ᶜAbdu) | 567 | 119,520 | 21,094 |
| طلال مداح (Ṭalāl Maddāḥ) | 523 | 93,836 | 13,926 |
| أصالة نصري (Aṣāla Naṣrī) | 371 | 80,318 | 13,890 |
| عمرو دياب (ᶜAmr Dīyāb) | 402 | 74,660 | 11,726 |
| عبد الله الرويشد (ᶜAbd Allāh al-Ruwaishid) | 413 | 68,320 | 12,062 |
| عبد المجيد عبد الله (ᶜAbd al-Majīd ᶜAbd Allāh) | 307 | 68,640 | 12,346 |
| كاظم الساهر (Kāẓim al-Sāhir) | 343 | 65,002 | 11,656 |
| راشد الماجد (Rāshid al-Mājid) | 360 | 65,271 | 10,352 |
| رابح صقر (Rābiḥ Ṣaqr) | 376 | 60,505 | 9,368 |

Table 5: Top lyric artists ranked by a balanced contribution score with equal weighting of songs, words, and verses.

| Poet | Poems | Words | Verses |
|---|---|---|---|
| ابن نباتة المصري (Ibn Nubāta al-Miṣrī) | 1,726 | 129,088 | 26,567 |
| الشريف الرضي (al-Sharīf al-Raḍī) | 677 | 147,974 | 31,953 |
| أبو العلاء المعري (Abū al-ᶜAlāʾ al-Maᶜarrī) | 1,609 | 113,018 | 23,576 |
| عبد الغني النابلسي (ᶜAbd al-Ghanī al-Nābulusī) | 962 | 120,823 | 26,508 |
| الشريف المرتضى (al-Sharīf al-Murtaḍā) | 587 | 130,799 | 27,561 |
| أحمد محرم (Aḥmad Muḥarram) | 437 | 130,615 | 27,399 |
| محمود درويش (Maḥmūd Darwīsh) | 470 | 117,286 | 23,357 |
| عبد الغفار الأخرس (ᶜAbd al-Ghaffār al-Akhras) | 378 | 112,462 | 23,669 |
| ابن الساعاتي (Ibn al-Sāᶜātī) | 519 | 102,802 | 21,225 |
| صفي الدين الحلّي (Ṣafiyy al-Dīn al-Ḥillī) | 898 | 86,857 | 18,236 |

Table 6: Top poets ranked by a balanced contribution score with equal weighting of poems, words, and verses.

# 6 Ethical and Legal Considerations

Tarab is intended for research use. The corpus contains text extracted from publicly accessible sources, including an openly released poetry dataset and lyric text from Kaggle[3], as well as material from the Habibi corpus (El-Haj, 2020). No audio, recordings, or musical compositions are included. Because lyrics and some modern poetic texts may be subject to copyright, we distribute Tarab with an explicit research-oriented usage statement and provide a takedown mechanism for rights holders. The release package is designed to support computational analysis of linguistic and stylistic patterns rather than to substitute access to original works. The dataset is publicly available on HuggingFace https://huggingface.co/datasets/drelhaj/Tarab.

# 7 Limitations and Future Work

While Tarab provides broad coverage of Arabic creative language, it is not without limitations. First, temporal metadata is coarse-grained for parts of the corpus, particularly for heritage poetry, where association with historical eras is used in place of precise dates. This limits fine-grained diachronic analysis at the year or decade level. Second, although Tarab captures substantial dialectal diversity, dialect labels are assigned at the verse or work level and do not account for intra-textual mixing or gradual register shifts within individual songs or poems. Similarly, stylistic categories such as musical style or performance tradition are maintained separately from the core schema and are not exhaustively annotated across the entire dataset. Finally, the corpus focuses on

verse-level textual representation and does not encode higher-level musical, prosodic, or performance features that are central to many forms of Arabic song. As a result, Tarab is best suited to linguistic and stylistic analysis rather than full multimodal or musicological study. Future work could address these limitations by enriching temporal metadata where feasible, expanding auxiliary annotations related to style and performance, and developing benchmark tasks that leverage Tarab's coverage of dialect, genre, and historical depth. Future work could also explore controlled extensions of the corpus that support evaluation of downstream NLP tasks such as dialect identification, authorship attribution, and stylistic transfer.

# 8 Conclusion

This paper introduces the Tarab corpus, a large-scale resource of Arabic creative language that brings together song lyrics and poetry across more than fourteen centuries, multiple genres, and a wide range of linguistic varieties, and is publicly available at https://huggingface.co/datasets/drelhaj/Tarab. By adopting the verse as a unified analytical unit and separating dialect, origin, and stylistic practice in its design, Tarab enables analyses that are difficult to support using existing Arabic corpora. Through detailed coverage statistics and lexical analyses, we showed that Tarab captures substantial dialectal diversity, clear genre differentiation, and a layered lexical structure spanning Classical Arabic, MSA, and regional varieties. The corpus also preserves cultural depth by representing both canonical poets and contemporary artists, providing a balanced view of Arabic creative production across time. Tarab is intended as a reusable resource for research in Arabic NLP, computational sociolinguistics, and digital humanities, supporting tasks such as dialect modelling, authorship analysis, stylistic variation, and representation learning. Future work could extend the corpus with richer temporal metadata, additional stylistic annotations, and task-specific benchmarks, further strengthening its role as a reference resource for Arabic creative language.

---

[3] https://www.kaggle.com/datasets/ahmedabelal/arabic-poetry

# References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*.

Maged S Al-Shaibani, Zaid Alyafeai, and Irfan Ahmad. 2020. Meter classification of arabic poems using deep bidirectional recurrent neural networks. *Pattern Recognition Letters*, 136:1–7.

Ali Alhazmi, Rohana Mahmud, Norisma Idris, Mohamed Elhag Mohamed Abo, and Christopher Ifeanyi Eke. 2024. Code-mixing unveiled: Enhancing the hate speech detection in arabic dialect tweets using machine learning models. *Plos one*, 19(7):e0305657.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mohamed Attia, Mohsen Rashwan, Ahmed Ragheb, Mohamed Al-Badrashiny, Husein Al-Basoumy, and Sherif Abdou. 2008. A compact arabic lexical semantics language resource based on the theory of semantic fields. In *International Conference on Natural Language Processing*, pages 65–76. Springer.

Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. 2016. Shamela: A large-scale historical arabic corpus. *arXiv preprint arXiv:1612.08989*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

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)*.

Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2017. Alkhalil morpho sys 2: A robust arabic morphosyntactic analyzer. *Journal of King Saud University-Computer and Information Sciences*, 29(2):141–146.

Abderrahim Boudlal, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane, MOAO Bebah, and Mostafa Shoul. 2010. Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In *International Arab conference on information technology*, pages 1–6. Elsevier Science Inc New York, NY.

Kareem Darwish. 2013. Arabizi detection and conversion to arabic. *arXiv preprint arXiv:1306.6755*.

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

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. *Culture*, 2:1–359.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association.

Mo El-Haj and Saad Ezzini. 2024. The multilingual corpus of world's constitutions (mcwc). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 57–66.

Nizar Habash, Houda Bouamor, and Kemal Oflazer. 2014. *A multidialectal parallel corpus of Arabic*. Carnegie Mellon University.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

AM Mutawa and Ayshah Alrumaih. 2025. Determining the meter of classical arabic poetry using deep learning: a performance analysis. *Frontiers in Artificial Intelligence*, 8:1523336.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference*, pages 7022–7032.

Abdenour Padillo-Saoud. 2019. *Open Islamicate Texts Initiative (OpenITI), 2016 [Reseña]*. Universidad Nacional de Educación a Distancia (España).

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition. Linguistic Data Consortium, Catalog Number LDC2011T11. ISBN 1-58563-595-2.

Faisal Qarah. 2024. Arapoembert: A pretrained language model for arabic poetry analysis. *arXiv preprint arXiv:2403.12392*.

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.

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

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the fourth arabic natural language processing workshop*, pages 175–182.