Murabaa: A comprehensive Resource Platform for Arabic Morphology

Karim Bouzoubaa[1], Driss Namly[2], Abdelhamid Jihad[3], Rachida Tajmout[1], Jamal Ezzouaine[4], Hakima Khamar[4]
[1]Mohammadia School of Engineers, Mohammed V University in Rabat, Morocco,
[2]Institute of African, Euro-Mediterranean and Ibero-American Studies, Mohammed V University in Rabat, Morocco,
[3]Institute of Studies and Research for Arabization, Mohammed V University in Rabat, Morocco,
[4]Faculty of Arts and Humanities, Mohammed V University, Morocco

Abstract

Arabic language faces technical and cultural challenges, including a lack of high-quality resources and the prevalence of regional dialects, which hinders the development of effective language processing systems. Therefore, the "Murabaa" platform was developed to transform Arabic linguistic knowledge into integrated digital resources. The platform aims to provide accurate digital content and promote the use of Arabic in various fields to bridge the gap between tradition and modernity by offering integrated linguistic resources for developing advanced research tools. The platform provides eight accurate dictionaries in the form of a website and a web application, contributing to the digitization of knowledge and its representation within the framework of standard lexical markup. In this study, we also conduct a quantitative comparison of the resources against similar ones to assess the quality of the linguistic knowledge they provide.

1 Introduction

The world is currently experiencing a major transformation driven by advances in artificial intelligence (AI). This progress is largely enabled by sophisticated models trained on vast linguistic corpora, which require extensive data and computational resources for training, development, and performance optimization. A substantial portion of these linguistic resources is derived from social media and other online communication platforms, where content is often noisy, informal, and frequently inaccurate. Within this landscape, Arabic occupies a particularly important position. It is one of the most widely spoken languages globally and serves as the official language in approximately 27 countries, with more than 330 million native speakers. In addition, Arabic is the liturgical language of nearly 1.6 billion Muslims worldwide and ranks among the top languages used on the internet, typically reported as the fourth most widely used online. Despite its historical depth and cultural richness, the Arabic language faces several pressing challenges. First, Arabic suffers from a notable lack of systematic efforts to digitize its core scientific and linguistic knowledge, which hinders the development of robust language technologies and applications. Existing initiatives often remain fragmented and limited in scope, without a comprehensive and integrated treatment of derivational, morphological, syntactic, and semantic levels. Second, the rapid spread of AI applications, particularly Large Language Models (LLMs), has led many users to develop the misconception that AI-generated outputs are inherently more reliable than established reference texts, even though such models are prone to issues such as hallucination. This unwarranted trust coincides with a declining status of standard and Classical Arabic among many speakers and the growing dominance of regional dialects and foreign languages. Together, these trends complicate efforts to preserve, standardize, and disseminate both traditional and contemporary Arabic linguistic knowledge, especially among younger generations. This Arabic's morphological richness, combined with its declining status, presents an urgent challenge that demands high-quality, standardized digital resources. In this context, there is an imperative need to convert the extensive body of Arabic linguistic knowledge — much of which is not only preserved in printed books and manuscripts, but also resides tacitly in the minds of linguists who may be unwilling or unable to effectively communicate and formalize it — into comprehensive, open-access digital formats. Such resources are crucial for supplying linguistically precise texts that can directly and indirectly enhance the training and performance of LLMs. Our overarching long-term goal is to systematically digitize and computationally represent the entirety of Arabic scientific linguistic knowledge, progressing from foundational to advanced levels. Given the ambi-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 398–407

tious scope of this endeavor, the present project concentrates specifically on Arabic morphology, while the treatment of syntax and semantics is deferred to later stages. To this end, we have developed the "Murabaa" [1] platform. This innovative digital infrastructure enriches Arabic linguistic resources by systematically collecting, organizing, updating, and expert-reviewing scientific knowledge on Arabic word structure. Murabaa delivers a comprehensive suite of resources — including dictionaries, grammars, and glossaries — that provide accurate, reliable data on morphological components. The remainder of this paper is organized as follows. Section 2 reviews related prior work on Arabic linguistic digitization and morphological resources. Section 3 presents a detailed overview of the Murabaa platform, covering its vision, linguistic infrastructure, and computational architecture. Section 4 introduces the evaluation and comparison of the Murabaa resources with existing linguistic resources. Section 5 discusses practical applications of the Murabaa resources in Natural Language Processing, including their role in the development of Arabic LLMs. Section 6 concludes with key takeaways and outlines directions for future research extending the Murabaa platform.

2 State of the art

To the best of our knowledge, no platform similar to Murabaa exists that aggregates all the resources needed to fully cover Arabic morphology. In contrast, available options are either standalone resources developed and shared with the community or open-source tools that incorporate such resources internally. Accordingly, in this literature review, we present the available resources that are comparable to our lexicons in Murabaa.

2.1 Standalone resources

Arabic stop words review identifies numerous standalone lists, the most notable ones are:

- Abu El-Khair (2017) : Three lists—a syntactic one (1,377 words), a corpus-based high-frequency list (235 words after manual review), and a combined version (1,529 words).
- Medhat et al. (2014): Corpus-based list of 1,061 words, derived from the top 200 frequent terms (validated as stop words), plus all affixed variants.
- Alajmi et al. (2012): Statistical extraction via frequency, mean/variance, and entropy; lists merged using Borda's rule to yield 200 words.
- Stop Words Project Balucha a (2014): GPL-licensed collection with 162 Arabic words, likely corpus-based (e.g., includes "billion," "force," "announced").
- Zerrouki Taha (2012): Rule-based generator yielding 13,016 inflected forms from a manually compiled lexicon across grammatical categories.

For broken plurals lexicons, we identified two resources:

- List of Arabic Broken Plurals (Attia et al. (2011)): An automatically extracted list of 2,562 broken plurals from a large contemporary corpus, including morphological patterns for both singular and plural forms.
- Elghamry (2010): A compilation of 7,194 Arabic nouns with their broken plural forms, automatically derived from the electronic edition of the Alwaseet Arabic-Arabic Dictionary.
- Neme Neme (2020) broken plural collections, which contains 10,000 entries.

Regarding the lemma and stem lexicons, we highlight the following resources:

- DIINAR Dichy et al. (2002): The DIctionnaire INformatisé de l'ARabe is a proprietary database containing approximately 119,693 lemmas, along with their stems and associated morpho-semantic features.
- Qabas Jarrar and Hammouda (2024): A lexicographic database that synthesizes data from 110 existing lexicons, covering about 58,000 lemmas (45,000 nominal, 12,500 verbal, and 473 functional words), all tagged with morphological features.

[1] Murabaa is the English transliteration of the original Arabic "مربع" which is the acronym of "منصة رقمية لبنية الكلمة العربية" which translates to "A digital platform for Arabic word structure"

- Arabic Morphological Dictionary Distributed by ELRA (2013): A resource with 4,912,749 stems, broken down into 3,374,852 nouns, 1,537,699 verbs, and 198 grammatical words.

2.2 Resources in tools

Regarding the resources made available through their tools, we distinguish between those that can be directly accessed via the tools and those embedded within the tools and therefore inaccessible to us, such as MADAMIRA (Pasha et al. (2014)) or most CAMeL tools (Obeid et al. (2020)). Among the resources that do provide access, the following can be mentioned:

- Khoja Shereen (2002) Arabic stemmer: Developed using a hybrid of statistical and rule-based methods, it incorporates lexicons for prefixes, suffixes, stop-words, roots, and patterns.
- Light10 (Larkey et al. (2007)): Among ten stem-based stemmers, Light10 stands out for its proven effectiveness in information retrieval on standard TREC datasets. Its clitics lexicon is hard-coded into the tool.
- ISRI (Taghva et al. (2005)): The Information Science Research Institute's (ISRI) Arabic stemmer is similar to Khoja's but without a root dictionary. All its resources (clitics, stop-words, and patterns) are hard-coded.
- Alkhalil Analyzer II (Boudchiche et al. (2017)): A context-independent Arabic morpho-syntactic analyzer using root-pattern matching, with lexicons for stop-words, roots, patterns, and clitics. Note that it relies on surface patterns, which include features such as tense, person, and clitics.
- FARASA Segmenter (Darwish and Mubarak (2016)): A morpheme segmentation tool powered by an SVM-rank model, which preprocesses using lookup lists of clitics, stop-words, roots, and patterns.

This review shows that Arabic resources have been studied briefly, yielding some published lists. However, these resources exhibit one or more of these limitations: absence of diacritics, dependence on particular corpora, insufficient coverage, inadequate interoperability, or omission of morphological characteristics.

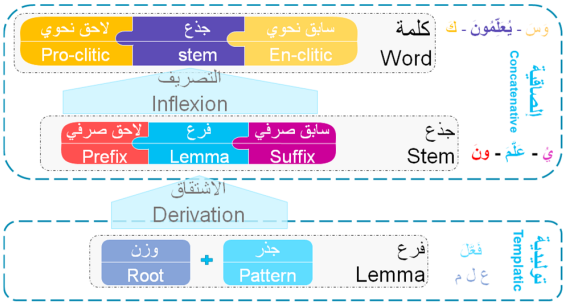

Figure 1: Incremental Word Formation Process

3 Murabaa Platform Project

3.1 Building methodology

The Murabaa platform is grounded in the classical Arabic grammatical framework, which categorizes words into nouns, verbs, and particles. This foundation enables comprehensive coverage of Arabic at lexical, morphological, derivational, and phonological levels. The platform implements a bottom-up methodology for Arabic word generation, starting from the alphabet as the atomic unit (Figure 1):

- **Root Generation**: Roots are assembled from alphabet letters guided by phonological constraints. This requires three lexicons: alphabet features, phonological rules for letter combinations, and Arabic roots with metadata.
- **Lemma Generation**: Grammatical patterns are applied to roots through nominal derivation and verbal inflection, producing lemmas and derived stems. Key outputs include a patterns lexicon and the "qalam" lexicon of lemmas/stems.
- **Special Lemmas**: Stop words and broken plurals represent non-templatic categories that deviate from standard Arabic morphology.

Overall, the approach yields nine interconnected lexicons: alphabet, phonological rules, roots, patterns, stems/lemmas (CALEM), functional words, broken plurals, clitics, and clitic rules. These enable precise computational simulation of Arabic morphology.

3.2 Computational infrastructure

Following the comprehensive representation of Arabic morphological knowledge across the nine lexicons described above, we computerized these

Figure 2: Excerpt from the Alphabet lexicon

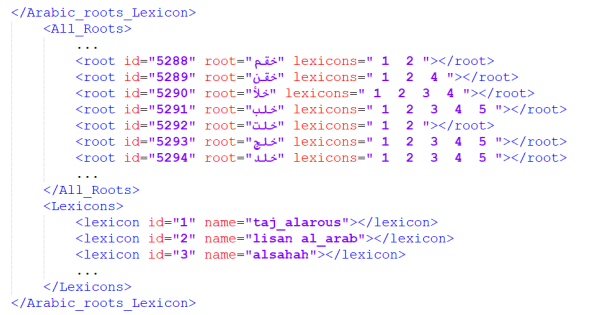

Figure 3: Excerpt from the lexicon of Roots

elements in a standardized format to ensure accessibility and usability for all researchers. To promote interoperability and reuse, the developed lexicons are encoded using standards such as the Lexical Markup Framework [2] (LMF) and made freely available [3] under a CC-BY-NC-ND license.

3.2.1 Alphabetical lexicon

The alphabetical lexicon encompasses 42 entries, comprising the 28 standard Arabic letters, five hamza forms, nine special letters, nine vowel markings, and three punctuation marks. For each entry, we annotated the following attributes: textual description, Unicode encoding, positional forms (initial, medial, final, isolated), and transliteration (Figure 2). Unlike prior implementations, where this lexicon is hardcoded directly into tools without a standalone structure, ours constitutes a distinct, queryable resource.

[2] http://www.lexicalmarkupframework.org/

[3] https://github.com/alelm-lab/Murabaa/

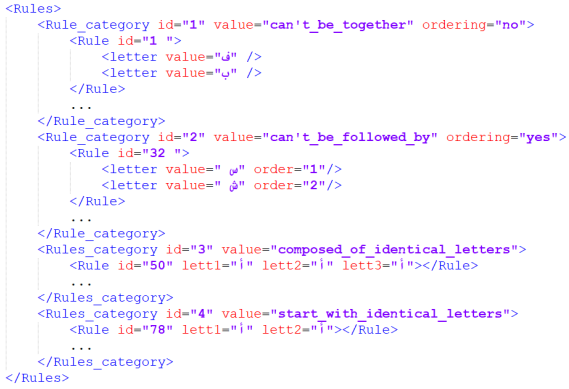

Figure 4: Excerpt from the lexicon of phonetic rules

3.2.2 Roots lexicon

Triliteral roots predominate in Arabic lexicography, accounting for 72% of entries. Accordingly, we prioritized compiling triliteral roots from the available lexicons, yielding a resource of 8,426 unique roots. Each root is annotated with metadata indicating the source lexicons in which it appears, as shown in Figure 3. For example, the root "خقم" (xqm) has the identification number 5288 and is tagged with lexicons 1 and 2, meaning it is found in the Taj al-Arus and Lisan al-Arab lexicons.

3.2.3 Phonetic rules lexicon

We compiled phonetic rules from the literature and structured them as an XML file for standardized representation. Each rule includes a unique identifier, a class label, and the incompatible character pairs it governs. Unlike conventional implementations, where these rules are hardcoded directly into processing tools without modularization, our approach yields a standalone, extensible lexicon. For example, in the excerpt in Figure 4, rule 2 prohibits the letter س (seen) from immediately preceding ش (sheen) in any Arabic root.

3.2.4 Patterns lexicon

The patterns lexicon is compiled through inductive analysis of Arabic morphological patterns documented in prior studies (Jamal Al-Zawain et al., 2023). This process yielded 378 entries, each annotated with a word class and illustrative example: 363 nominal patterns and 15 verbal patterns. For instance, the pattern "أَفْعَل" (>afoEal; ID: 1) functions as either a superlative noun (class 2g) or a derived adjective (class 2h).

3.2.5 Clitics lexicon

The clitics lexicon catalogs 12 proclitics and 14 enclitics as atomic units. Complex forms arise by combining these units in specific orders via defined association rules—for instance, the proclitic "أَسَ" (>asa) assembles from "أَ" (>a) and "سَ" (sa)—yielding approximately 94 proclitics and 73 enclitics. Conventional Arabic NLP tools typically embed these clitics in pre-compounded forms or hardcode association rules, obscuring components. To address this and ensure modularity compliant with linguistic resource standards, we formalized the lexicon in an XSD schema.

3.2.6 Stop words lexicon

This lexicon aggregates Arabic stop words, classified into three subgroups based on distinct grammatical properties: 69 particles, 180 special nouns, and 66 special verbs. Each entry includes morphological and syntactic annotations; for example, "لكن" (lkn) functions as either an introductory contrast particle or a contrastive conjunction. In a two-step process, we first compiled these 515 base (simple stop words) forms. We then systematically affixed possible clitics to generate compound stop words, resulting in 17,153 total entries. This substantially exceeds prior resources, which max out at 1,529 entries (Abu El-Khair, 2006).

3.2.7 Broken plurals lexicon

We constructed this lexicon by systematically extrapolating broken plural forms from legacy Arabic lexicons (Ouamer et al. (2022)), yielding 12,249 unique entries. This surpasses the largest prior collections, which do not exceed 10,000 entries (Neme and Laporte, 2013). For instance, the singular noun جَمَل (camel) maps to eight broken plurals, including أَجْمَال، جِمَال and جَمَائِل.

3.2.8 Calem lexicon

The CALEM lexicon is built from a database of Arabic verbs (24,171 entries derived from roots) (El Jihad et al. (2018)). We generated conjugated verb forms, then applied derivational patterns to yield derived nouns; the lexicon is further augmented with non-derived nouns, including proper nouns. This comprehensive approach produced 166,963 lemmas across 7,133,106 stems, far exceeding prior resources with only 122,000 entries (Ramzi et al., 2004). For example, the noun "كاتب" (writer) derives from the root "ك ت ب" (to write) via the "فاعل" (fAEl) pattern and includes subentries like "كاتبان" (two writers; Figure 5).

```
<Lexicon>
    <feat att="language" val="arab" />
    <LexicalEntry id='كَاتِبٌ' >
        <feat att='partOfSpeech' val='n' />
        <Lemma>
            <feat att='writtenForm' val='كَاتِبٌ' />
            <feat att='scheme' val='فَاعِلٌ' />
        </Lemma>
        <WordForm>
            <feat att='wittenForm' val='كَاتِبٌ' />
            <feat att='prefix' val='#'/>
            <feat att='suffix' val='#'/>
        </WordForm>
        <WordForm>
            <feat att='wittenForm' val='كَاتِبَانِ' />
            <feat att='prefix' val='#'/>
            <feat att='suffix' val='ان'/>
        </WordForm>
        ...
        <RelatedForm targets='Racine' >
            <feat att='type' val='كتب' />
        </RelatedForm>
    </LexicalEntry>
</Lexicon>
```

Figure 5: Excerpt from CALEM lexicon

3.3 The Murabaa platform

The platform is deployed as both a static website and an interactive web application. The website (Figure 6) provides:

- A concise project overview.
- A demonstrative video covering the homepage and resource navigation.
- Testimonials underscoring the project's contributions.
- Comprehensive bibliographic references.
- Direct download links to all lexicon files encoding Arabic morphological knowledge.

Activating the "Demo" button redirects users to the web application for resource exploration. The application supports Arabic and English interfaces, with a right-side menu listing all available resources. Selecting a menu option loads the corresponding resource in the central panel. Users browse paginated content bidirectionally or jump to specific pages, with key navigation enhancements including a "Filter" input for advanced search criteria (dynamically updating displayed entries to match) and single-click selection of any entry (revealing detailed attributes in the left sidebar).

4 Resources evaluation and comparison

Our lexicons were developed in alignment with Arabic language structures, adhering fully to interoperability guidelines and validated by our team of

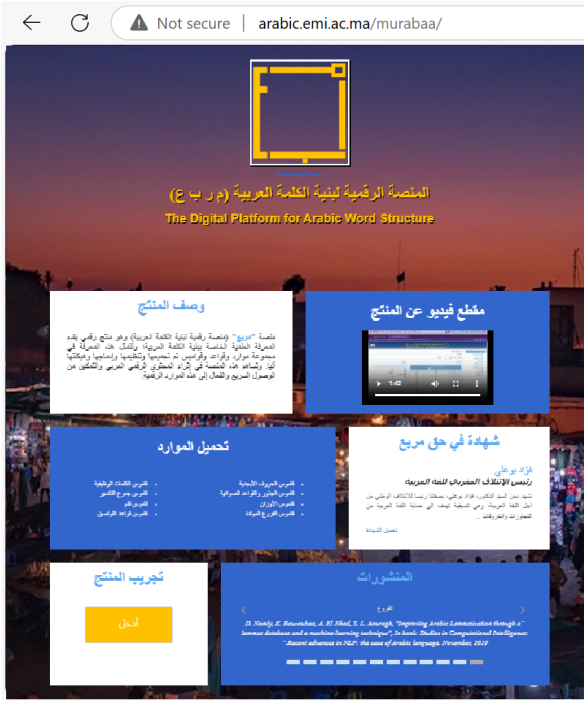

Figure 6: "Murabaa" website

linguists. To demonstrate their comprehensiveness, we conduct a rigorous evaluation of their overall quality.

To control the lexicon's quality we assess both qualitative and quantitative issues. The qualitative judgment is done through the lexicons evaluation, while the quantitative appraisal is carried out across the lexicons comparison.

4.1 The qualitative evaluation

Lexicon evaluation typically employs one of two approaches: comparison against a reference lexicon using standard metrics such as precision, recall, and F-score, or measurement of lexical coverage over a large annotated corpus. To the best of our knowledge, no suitable reference lexicon exists for Arabic resources. We therefore evaluate our lexicons using lexical coverage on large annotated corpora. The evaluation drew on three corpora, selected based on availability, morphological annotation quality, and size:

- Arabic-PADT UD Corpus (Smrz et al. (2008)): A large newswire collection in Modern Standard Arabic, comprising 189,860 morphologically and syntactically annotated words.
- Al-Mus'haf Corpus (AMC) (Zeroual and Lakhouaja (2016)) : 78,250 morphologically annotated words from the Holy Quran.
- NEMLAR Written Corpus (NC) (Kadim and LAZREK (2025)): Approximately 500,000 annotated words across 13 categories of Arabic text.

The goal of our evaluation is to measure the proportion of words from the corpora that appear in our lexicons. Lexical coverage is typically quantified using two metrics:

- Vocabulary coverage (VC): Proportion of unique (distinct) words from the corpus covered by the lexicon (duplicates counted once).
- Real text coverage (RC): Proportion of all tokens in the corpus covered by the lexicon (duplicates counted separately).

Vocabulary coverage often yields lower rates on large corpora, as frequent words dominate and are readily covered, while rare ones reduce the overall score. We computed both metrics for a comprehensive assessment. As shown in Table 1, our lexicons achieved perfect coverage for the alphabet, root, pattern, stop-word, plural, and clitic features across all corpora, with coverage exceeding 99% for both real text and vocabulary. We note that PAT lacks results for root and pattern because the corpus does not include these tags. For stem and lemma features, coverage is very high on the AMC and NC corpora (97.26%–99.89%). On the PADT, RC remained promising at approximately 95%, though VC dropped to 83.28% for stems and 78.63% for lemmas—noticeably lower than on the other corpora. The manual analysis shows that most of the non-covered lemmas are named entities, and the variation between VC and RC for the PADT corpus is mainly due to the presence of named entities with a large number and low frequency.

4.2 The quantitative comparison

Comparison involves evaluating quantities and/or features across at least two objects to identify similarities or differences. In our approach to assess the quality, the quantitative comparison entails evaluating both the size and features of our static lexicons against available benchmarks.

4.2.1 Roots

In this comparison, we assess the size of our roots lexicon against entries from three lists in the tools

		PADT	AMC	NC
VC (%)	Alphabet	100	100	100
	Root	-	99.82	99.68
	Pattern	-	100	100
	SW	100	100	100
	Plural	99.02	100	100
	Clitic	100	100	100
	Stem	83.28	97.26	NA
	Lemma	78.63	99.32	98.42
RC (%)	Alphabet	100	100	100
	Root	-	99.99	99.99
	Pattern	-	100	100
	SW	100	100	100
	Plural	99.95	100	100
	Clitic	100	100	100
	Stem	94.47	99.19	NA
	Lemma	95.11	99.89	99.34

Table 1: Evaluation results.

reviewed in the state-of-the-art section. The results show Murabaa as the largest roots lexicon: Alkhalil2 contains 5,774 roots, FARASA has 6,858, Khoja has 3,823, and Murabaa leads with 8,426. These differences in lexicon scale can significantly affect morphological analysis coverage and performance.

4.2.2 Patterns

This analysis sought to measure the sizes of lexicons among the assessed lexicons. Murabaa possesses the most extensive patterns lexicon. ISRI has 44 patterns, FARASA has 125, Khoja has 45, and Murabaa has 378.

4.2.3 Stop words

In this comparison, we evaluate the number of lexical entries and features in our stop-words lexicon against the five standalone lists mentioned earlier in Section 2. We compare the lexicons based on four features: i- Diacritized: Whether the list includes diacritized entries. ii- Rules-based: Examination of the compilation technique.iii- Classified: Whether the list is categorized. iv- Cliticized forms: Whether the list includes cliticized forms of the stop-words.

The comparison results show that the Murabaa stop-words lexicon substantially outperforms all compared lists with 67,153 entries. Moreover, the feature comparison scores indicate that Murabaa is the most comprehensive, as it is diacritized, rule-based in compilation, classified, and includes cliticized forms of the stop-words. The Arabic stop-words and Abu El-Khair lexicons rank second and third, respectively.

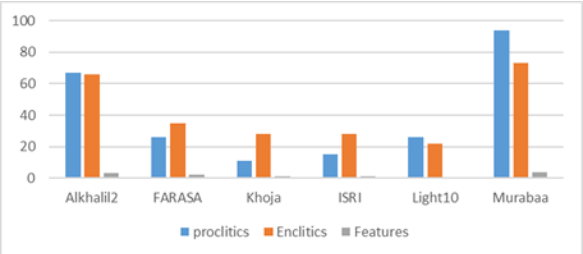

Figure 7: Comparison results for clitics lexicon

4.2.4 Broken plurals

A comparison of broken plural lexicons shows that our lexicon surpasses the others: Attia has 2,562 entries, Elghamry has 7,194, Neme has 10,000, and Murabaa has 12,249.

4.2.5 Clitics

This comparison aimed to quantify lexicon sizes (entry extent) and evaluate feature presence across lexicons. Specifically, we assessed the six available lexicons presented in the state of the art section with the following features: i- Atomic contents: Presence of atomic clitics (vs. only combined forms). ii- Explanation of atomic contents: Inclusion of descriptions for atomic clitics. iii- Constructors modeling: Coverage of rules governing combined clitic formation. iv- Association constraints modeling: Inclusion of compatible word types for clitics. As shown in Figure 7, our clitic lexicon contains the largest number of entries for both proclitics and enclitics, surpassing Alkhalil2 and FARASA. It also achieved the highest feature coverage among the compared lexicons

4.2.6 CALEM

CALEM is compared solely to the DIINAR lexicon, as the other similar resource "Arabic Morphological Dictionary" is not freely available and lacks published statistical data. In contrast, although DIINAR itself is not free, its quantitative figures have been published (Abbès, Dichy, and Hassoun 2004). The quantitative comparison reveals that CALEM has more entries for lemmas (166,963 against 121,522) and non-verbal stems (4,847,427 against 4,714,222), while DIINAR leads in verbal stems (3,060,716 against 2,464,239). This outcome seems puzzling: since stems are derived from lemmas and CALEM has more lemmas than DIINAR,

one would expect CALEM's stems to outperform DIINAR's, especially given the well-defined verbal inflection paradigms established by linguists. Unfortunately, DIINAR's lack of availability prevents us from investigating the main cause of this discrepancy.

5 Applications of Murabaa Resources in Natural Language Processing

The Murabaa platform provides a unified, richly annotated, and standardized set of lexical and morphological resources that enable a wide spectrum of NLP applications. The explicit structuring of Arabic linguistic knowledge—across roots, patterns, stems, clitics, stop words, phonological constraints, and broken plural forms—offers capabilities that are not commonly available in existing Arabic NLP tools. In this section, we illustrate how these resources can be effectively exploited in computational contexts.

First, it is important to note that many of the murabaa resources have been effectively leveraged in the development of various systems—not only for individual tools such as the Arabic learning application [4] or small Arabic games [5], but also for a comprehensive Arabic NLP infrastructure called Safar (Bouzoubaa et al. (2021)). This infrastructure includes more than fifty tools ranging from transliteration, tokenization, and POS tagging to more advanced components such as morphological analysis and summarization. The integration of our resources within this ecosystem has demonstrated their effectiveness, largely due to their high quality and grounding in standardized and well recognized benchmarking practices.

Second, Murabaa's nine interconnected lexicons provide the building blocks for determin- istic and learning-based morphological tools. For instance, Roots + Patterns + Stems (CALEM) re- sources enable large-scale morphological decompo- sition, useful for lemmatization, stemming with lin- guistic guarantees and root-pattern alignment tasks. Another example is the exploit of Clitics lexicon + Clitic attachment rules support robust tokenization and de-cliticization, overcoming the fragmentation and inconsistency seen in third-party Arabic segmenters. In a word, such structured resources enable deterministic morphological analyzers, hybrid analyzers, and even supervised neural models that benefit from explicit morphological constraints.

Finally, Murabaa resources can help build cleaner pipelines and more linguistically grounded prepro- cessing modules. Indeed, many NLP systems for Arabic—such as dependency parsers, NER models, and machine translation pipelines—suffer from ambiguities caused by concatenative morphology, clitic fusion, absence of short vowels, or homography across lemmas. The Murabaa platform mitigates these issues. For example, the phonological rules lexicon prevents generation of invalid roots or stems, reducing noise in models that use synthetic data. Also, the Stop-word lexicon (simple + compound) enables high-coverage filtering for topic modeling, retrieval, and text classification. One of the most strategic uses of the Murabaa platform lies in the development of Arabic LLMs, which currently lag behind English and other high-resource languages due to limited high-quality linguistic datasets. Murabaa can contribute to LLM design at several levels. This is the case for Improving Tokenization and Vocabulary Construction. Current tokenizers (SentencePiece, BPE, WordPiece) treat Arabic as opaque, often splitting roots from patterns, clitics from stems, or templatic morphology into arbitrary sub-tokens. Murabaa resources allow developers to build morphologically aware tokenizers, enforce valid clitic + stem boundaries, design root-pattern consistent subword vocabularies, reduce vocabulary sparsity for verb and noun inflection families, and avoid allocating tokens to invalid or non-existent stems. This leads to smaller vocabularies, fewer unknown tokens, and better generalization across word families. Murabaa lexicons can be used to define Linguistic Constraints During Training and Evaluation. It can be used for morphological evaluation benchmarks (e.g., verb conjugation accuracy, plural prediction), diagnostic tasks for probing LLMs, and constraints during decoding (e.g., prohibiting illegal clitic combinations using the clitic rules lexicon). This enables both safer generation and more interpretable model behavior. Also, because Murabaa expresses knowledge in structured LMF format, it can serve as a source for knowledge-grounded pretraining (e.g., injecting stem–root links), embedding alignment tasks, and training lexical or morphological adapters. Overall, the Murabaa platform provides the largest standardized morphological resource suite for Arabic, enables morphologically informed preprocessing and

[4] https://arabic.emi.ac.ma:8080/iLearnArabic/
[5] https://arabic.emi.ac.ma/games/

hybrid systems, gives LLM developers access to high-quality linguistic constraints, synthetic data, and specialized vocabularies, and bridges the gap between traditional Arabic linguistic competence and modern AI methodologies.

6 Conclusion

The digital platform "Murabaa" serves as a fundamental linguistic reference for Arabic language research, contributing to the development of a digital morphological framework for Arabic letters, roots, and patterns, based on digital resources specifically designed for this purpose. Arabic language computing cannot become a reality without the necessary resources and tools to enrich digital data with accurate linguistic information that adheres to the rules and regulations of the Arabic language.

The research team aspires for "Murabaa" to become a standard reference platform for specialists and researchers in Arabic language computing. It is a platform that ushers in a new era of reliable digital linguistic knowledge. In doing so, "Murabaa" adds scientific value to the field of computational linguistics by providing search engines with comprehensive, integrated, and systematically structured data.

Limitations

The evaluation of quality control and validation for the created lexicon entries is conducted through coverage, which is predominantly based on the size of the lexicon, rather than on functional performance assessed through precision-focused audits or reported manual spot-check statistics. This situation arises primarily from the lack of comparable resources. Furthermore, we can illustrate the influence of the developed lexicons on NLP tasks within downstream NLP pipelines; however, due to time constraints, we would rather address this in other works.

Acknowledgments

In the preparation of this paper, we acknowledge the use of AI to enhance the clarity and coherence of the English language and ensure it adheres to the standards expected, but it did not contribute to the generation of ideas.

References

Amal Alajmi, E Mostafa Saad, and RR Darwish. 2012. Toward an arabic stop-words list generation. *International Journal of Computer Applications*, 46(8):8–13.

Mohammed Attia, Pavel Pecina, Lamia Tounsi, Antonio Toral, and Josef van Genabith. 2011. Lexical profiling for arabic. *Proceedings of eLex*, pages 23–33.

Balucha a. 2014. Stop words project. Accessed: January 10, 2026.

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

Karim Bouzoubaa, Younes Jaafar, Driss Namly, Ridouane Tachicart, Rachida Tajmout, Hakima Khamar, Hamid Jaafar, Lhoussain Aouragh, and Abdellah Yousfi. 2021. A description and demonstration of safar framework. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 127–134.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074.

Joseph Dichy, Abdelfattah Braham, Salem Ghazali, and Mohamed Hassoun. 2002. La base de connaissances linguistiques diinar. 1 (dictionnaire informatisé de l'arabe, version 1). In *Proceedings of the International Symposium on The Processing of Arabic, Tunis (La Manouba University)*, pages 18–20.

Distributed by ELRA. 2013. Arabic morphological dictionary. Accessed: January 10, 2026.

Abdelhamid El Jihad, Driss Namly, fettah Hamdani, and Karim Bouzoubaa. 2018. The development of a standard morpho-syntactic lexicon for arabic nlp. In *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, pages 1–5.

Ibrahim Abu El-Khair. 2017. Effects of stop words elimination for arabic information retrieval: a comparative study. *arXiv preprint arXiv:1702.01925*.

K Elghamry. 2010. A lexical-syntactic solution to the problem of broken plural in arabic. *G. URT G*.

Mustafa Jarrar and Tymaa Hasanain Hammouda. 2024. Qabas: An open-source arabic lexicographic database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13363–13370.

Ayoub Kadim and AZZEDDINE LAZREK. 2025. Nemlar corpus improvement for arabic natural language processing. *Journal of Theoretical and Applied Information Technology*, 103(4).

Khoja Shereen. 2002. Khoja stemmer. Accessed: January 10, 2026.

Leah S Larkey, Lisa Ballesteros, and Margaret E Connell. 2007. Light stemming for arabic information retrieval. In *Arabic computational morphology: knowledge-based and empirical methods*, pages 221–243. Springer.

Walaa Medhat, Ahmed H Yousef, and Hoda Korashy. 2014. Corpora preparation and stopword list generation for arabic data in social network. *arXiv preprint arXiv:1410.1135*.

Alexis Neme. 2020. *An arabic language resource for computational morphology based on the semitic model*. Ph.D. thesis, Université Paris-Est.

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.

Mariame Ouamer, Rachida Tajmout, and Karim Bouzoubaa. 2022. Arabic broken plural model based on the broken pattern. In *Digital Technologies and Applications*, pages 22–31, Cham. Springer International Publishing.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec*, 2014, pages 1094–1101.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Otakar Smrz, Viktor Bielickỳ, Iveta Kourilová, Jakub Krácmar, Jan Hajic, and Petr Zemánek. 2008. Prague arabic dependency treebank: A word on the million words. In *Proceedings of the workshop on Arabic and local languages (LREC 2008)*, pages 16–23.

Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, volume 1, pages 152–157. IEEE.

Imad Zeroual and Abdelhak Lakhouaja. 2016. A new quranic corpus rich in morphosyntactical information. *International Journal of Speech Technology*, 19(2):339–346.

Zerrouki Taha. 2012. Arabic stopwords. Accessed: January 10, 2026.