A Comprehensive Full-Form Lexicon for Arabic NLP and Speech Technology

Jack Halpern & Yannis Haralambous
CJK Dictionary Institute, Inc. / IMT Atlantique & UMR CNRS 6285 Lab-STICC
Niiza, Saitama, Japan / Brest, France
jack@cjki.org / yannis.haralambous@imt-atlantique.fr

Abstract

Natural Language Processing (NLP) applications require morphological data with precise grammatical attributes, while speech technology requires abundant phonemic and phonetic data. This presents a challenge for Arabic due to its abundant morphological, orthographic, and phonemic ambiguity in both MSA and its various dialects. Existing systems struggle with incomplete and unstructured web data, leading to suboptimal performance in both morphological analysis and speech applications. This paper presents ArabLEX, a *full-form lexicon* (includes all wordforms, i.e., fully inflected/cliticized members of a lexeme class) that addresses these issues by providing a large-scale database designed to enhance NLP accuracy. It comprises approximately 570 million entries with fully inflected forms and detailed morphological, phonetic, and orthographic attributes. ArabLEX serves as a foundational framework for developing comprehensive Arabic lexical resources for NLP, particularly for speech technology, as well as dialect databases.

Keywords: Arabic language, MSA, Arabic writing system, graphemic underrepresentation, full-form lexicon, NLP, speech technology

1. Introduction

1.1. What is a Full-Form Lexicon

Traditionally, dictionary headwords consist of canonical forms (lemmata). A *full-form lexicon* explicitly includes all wordforms of a language, i.e., fully inflected, conjugated, declined, or cliticized ("inflected" for short) members of a lexeme class, rather than just the lemmata. For example, the English lexemes *eat* and *boy* have the members *eat*, *eats*, *eating*, *eaten*, *ate* and *boy*, *boys*, *boy's*, *boys'* respectively. For highly inflected and agglutinative languages, the abundance of combinatorics (stem, affixes, clitics) can result in full-form lexicons with hundreds of millions of entries.

The morphology of Arabic[1] is *templatic*, in the sense that it is based on *roots* and *patterns* (Ryding, 2005). This means that we are not just dealing with stems and affixes as in many Indo-European languages, but with tri- or quadriliteral consonantal roots with infixes, prefixes, suffixes, and circumfixes. The morphological generative principle inherent to Arabic morphology is omnipresent and even applies to loanwords (Gadelli, 2015); it can thus be considered as an innate property of Arabic. Arabic full-form lexicons are naturally of a large size since they are bound to cover a very large number of Arabic root + pattern + clitic combinations (so that all grammatical wordforms are available to the user), both in the oral modality (represented by phonetic and phonemic transcriptions) and in the written modality of the Arabic language (represented by graphemic and graphetic sequences, cf. Meletis and Dürscheid (2022)).

[1] Arabic refers to MSA, the official language of 380M people, but (practically) no one's mother tongue (Haugen, 1972; Mejdell, 2014), as well as its dialects.

1.2. State of the Art

Several tools have been developed for morphological analysis, tokenization, generation of inflected and conjugated forms, POS tagging, and disambiguation of the Arabic language. We refer to such tools as *morphological engines*. Among the most popular tools we find AlKhalil (Boudchiche et al., 2017), MADA (Habash et al., 2009), BAMA (Buckwalter, 2002), PATB (Penn Arabic Treebank) (Maamouri et al., 2004), FARASA (Abdelali et al., 2016), MADAMIRA (Pasha et al., 2014), Elixir_FM (Smrž, 2007), CALIMA Star (Taji et al., 2018), the most recent Camel Morph MSA (Khairallah et al., 2024), a highly ambitious large-scale resource, and ArabLEX, which we present in this paper.

The processing performed by morphological engines is supported by lexical databases, such as tables for stems, clitics, and affixes (Sawalha, 2011). Despite their high performance (Taji et al., 2018), most of these tools (with the notable exception of Camel Morph) have shortcomings, including inconsistency, ignorance of lexical rationality, and the lack of phonological attributes. Still, the goal of these tools is to perform computational tasks, such as tokenization and disambiguation, rather than to serve as comprehensive lexicons enumerating all possible wordforms. A notable exception worth mentioning is the Arabic full-form lexicon and Finite State Transducer (FST) project by Soudi and Eisele

(2004), which offers a targeted solution focused on Arabic morphology.

1.3. Full-form vs. Generation

Morphological generators (MG), a subclass of morphological engines, are algorithms that generate wordforms on demand *in real time*, whereas full form lexicons (FFL) are *static databases.* However, FFLs also have a generation phase based on rules and templates. The difference is that in the case of FFLs, generation is done at the resource creation phase (see §4), not in real time. However, if an FFL is provided with a user interface, it can behave exactly like an MG; that is, do both analysis and generation.

Among the advantages of FFLs is (1) their data can be stored in a simple, processing-friendly structure and format (TSV) which makes it easy to import into SQL databases; (2) the simple text format enables direct access, search and manipulation without special tools; (3) the data is easy to integrate into LLMs for improving their ability to handle complex morphology; (4) they can help LLMs improve tokenization, embed morphological features, and support retrieval-augmented generation; and (5) during inference, FFLs ensure morphological validity and agreement and reduce instances of out-of-vocabulary (OOV). Among the disadvantages of FFLs is (1) their very large file size (e.g., ArabLEX is over 59GB); and (2) updates of the rules used as part of the generation step are computationally intensive.

Among the advantages of MGs is (1) their relatively compact file size (e.g., Camel Morph is 80MB); (2) their rules are easy to update and, in the case of rule revision, there is no need to regenerate; and (3) with some tools, the morphological rules used for generation may be accessible to the user in an explicit form. Finally, among the disadvantages of MGs is (1) their raw data is stored in a complex, processing-unfriendly structure; (2) direct access typically requires special tools; (3) SQL integration is cumbersome; and (4) they cannot be easily integrated into LLMs.

1.4. Arabic Full-Form Lexicons and LLMs

A full-form lexical resource can enhance an LLM's ability to generate morphologically accurate Arabic text and reduce OOV issues, which is particularly important given Arabic's rich inflectional and derivational morphology, where a single root can produce up to 5,000 or more distinct wordforms.

Integration approaches include morphologically-aware tokenization during pre-training (Zhu et al., 2024), where lexicons guide segmentation strategies to preserve meaningful linguistic units rather than relying solely on frequency-based methods like BPE (Sennrich et al., 2016). Architecturally, lexicons can be integrated through retrieval-augmented generation (RAG) or as structured embeddings encoding morphological features—part-of-speech tags, inflectional categories, and derivational patterns—enabling models to learn systematic relationships between wordforms sharing the same root (a natural extension of (Cotterell and Schütze, 2015)).

During inference, lexicons serve as constraint mechanisms for morphological validity, particularly for agreement phenomena. This can be implemented through a modified beam search that penalizes invalid candidates or post-generation correction modules (Lu et al., 2021). Additionally, comprehensive lexical coverage addresses Arabic NLP's data sparsity problem by providing explicit mappings between surface forms and morphological analyses, enabling models to handle rare but valid inflectional variants that may be absent from training corpora.

1.5. Conventions

The phonemic transcriptions in this paper are italicized and given in the CARS system (Halpern, 2009a). Transliterations are given in the Buckwalter transliteration system (Buckwalter, 2002) and enclosed in forward-slashes. Linguistic terminology and, in particular, definitions of the terms *graphemic, graphetic*, *writing system*, *script*, etc., are based on Haralambous (2024).

2. Writing System Ambiguities

2.1. Templatic Morphology

In templatic morphology, inflection is performed by changing the vowel + consonant patterns by affixation and cliticization. By cliticization, we refer to the fact that not only can words be declined and conjugated ("inflected" for short), but they can also take clitics. For example, adding the proclitics *wa* 'and', *li* 'to', and the enclitic *ātíhimạ* to the stem *kắtib* 'writer' yields the complex form *walikạtibātíhimạ* (وَلِكَاتِبَاتِهِمَا) 'and to their (two) (female) writers'. This type of combinatorics yields a vast number of wordforms. For example, the full paradigms for كَاتِبٌ *kātibun* 'writer' and كَتَبَ *kataba* 'write' reach about 5,660 and 6,900 forms, respectively. Templatic morphology guarantees that the combinatorially obtained forms are, in their vast majority, phonemically distinct, making it impossible to have homophones obtained from different roots and patterns.

2.2. Layered Representation

This is not the case in the written modality, due to the *layered nature* of the Arabic writing system. Indeed, the Arabic script uses seven layers of signs, expanding from the skeleton (*rasm*) to points (*nuqta*), gemination (*shadda*), short vowels and nunation (*tashkil* and *tanwiin*), as well as three other layers specific to the Quran (Osborn, 2017).

In a formal context, the MSA writing system[2] uses the first four layers—a text using the first four layers is commonly called *fully vocalized*, even though there is more than vowels in the fourth layer. Such a text achieves an almost one-to-one mapping between phonemes and graphemes, so that the many *phonemically* distinct forms resulting from the fertility of inflection, cliticization, and derivation are also *graphemically* distinct. But in everyday use of the language, in media and the private sphere, only the first three layers (sometimes even only two, when the gemination sign *shadda* is omitted) are used, resulting in texts that are *graphemically underrepresented* ("orthographically ambiguous").

2.3. Orthographic Ambiguity

We have two kinds of ambiguity when going from the written to the oral modality.

Relatively rare cases where even the fully layered writing system is unable to accurately represent the grapheme-phoneme mapping, for example in (a) the *ʾalif alfaaSila* (otiose *alif*) (Ryding, 2005), where the *alif* is not pronounced and exists solely as a marker of the 3rd plural masculine of the perfect active (e.g., كتبوا being phonemically realized as *katabu*), (b) vowel neutralization sometimes being lexically determined and thus unpredictable from the orthography, e.g., في القاهرة 'in Cairo', the preposition في is pronounced *fi*, not *fii*; also نا *naa* is written as a long vowel in أنا but is shortened to *na* when uttered (see Halpern (2009c)), etc.

The vast majority of ambiguous cases resulting from the lack of the fourth (and sometimes also of the third) layer, for example, (a) the absence of short vowels (e.g., كاتب represents the seven wordforms *kātib*, *kātibun*, *kātibin*, *kātaba*, *kātibi*, *kātiba*, *kātibu*), or (b) the omission of *shadda* indicating consonant gemination, e.g., محمد (diacritized مُحَمَّد), which provides no clues that the م is geminated, (c) the representation of long *ā* by ا as in سوريا or by آ as in آسيا, but some bare alifs representing *tanwiin* rather than long *ā*, as in شكرا *shukran*, etc.

[2] As opposed to other Arabic-script languages like Persian, Ottoman, or Uyghur, which are not based on based on the abjad system of morphological inference based on consonants and long vowels (Dichy, 2017; Haralambous, 2021).

3. ArabLEX

3.1. Introduction

To avoid the ambiguity inherent in the Arabic writing system presented in the previous section, it is important to have a large database of wordforms. ArabLEX is a full-form lexicon that offers exactly that, since it is designed for maximal lexical and morphological coverage.

3.1.1. Scope and Coverage

The first release of ArabLEX in 2021 covered about 530 million entries for general vocabulary and proper nouns. In the second release, another 40 million entries have been added, expanding the number of entries to 570 million data lines. ArabLEX consists of four main modules: DAG (Arabic General Vocabulary, 83M entries, The CJK Dictionary Institute (2019b)), DAN (Arabic Names, 218M entries, The CJK Dictionary Institute (2019a)), DAF (Arabic Foreign Names, 226M entries, The CJK Dictionary Institute (2019d)), and DAP (Arabic Place Names, 6M entries,The CJK Dictionary Institute (2019c)). The number of data fields (up to 25) varies with the database module (see §3.2). These provide detailed grammatical, phonological, morphological, and orthographic attributes (Halpern, 2020).

Besides MSA, ArabLEX has also served as a foundational framework to develop a series of full-form lexicons for the major Arabic dialects called DiaLEX (The CJK Dictionary Institute, 2025a), containing about 150 million entries based on the same methodology used for ArabLEX.

3.1.2. Target Users

ArabLEX, to our knowledge the largest full-form Arabic lexical resource available, serves as a valuable resource for linguistic researchers conducting detailed analyses of Arabic morphology. It is also aimed at developers of Arabic NLP software, such as morphological analyzers, machine translation, automatic speech recognition (ASR), text-to-speech (TTS), and other NLP language-processing tools. For academic research, the ArabLEX core module(s) or subsets are available free of charge; for commercial development, it is available for licensing or through ELRA.

3.2. Database Structure and Attributes

ArabLEX consists of a set of flat text files in TSV format, encoded in UTF-8. The number of fields (maximally 25) is file-dependent. Fields can be classified into categories such as phonetic (IPA), phonemic (CARS), morphological (STEM, CASE),

grammatical (GEN, NUM), semantic (ROOT, RAT, i.e., rationality), and others.

The simplicity of the format makes it easy to access, process, and import data into database systems. Queries to ArabLEX can be submitted through a user interface or from the command line with the appropriate parameters. It can also serve as a foundational database for morphological engines.

In the following, we describe ArabLEX's various attributes.

3.2.1. Grammatical Attributes

The grammatical attributes of ArabLEX are useful for morphological analysis, orthographic disambiguation, POS tagging, semantic analysis, and more. These include codes for gender, number, case endings, and person, as well as the stem, definiteness property (binary), root, and lemma.

Data field	Value	Data field	Value
Full-form	وَلِكَاتِبِكُمَا	Number	D (dual)
Lemma	كَاتِبٌ	Person	2 (second)
Stem	كَاتِب	Definiteness	D (definite)
Gender	C (common)	Root	ك-ت-ب
Case	GEN (genitive)		

Table 1: Grammatical attributes

3.2.2. Phonological Attributes

The phonemic and phonetic transcriptions are useful for improving speech technology, be it TTS or ASR (Tahon et al., 2016; Feng et al., 2023). These include fully diacritized Arabic graphemic strings, as well as phonemic and phonetic transcriptions, including word stress and vowel neutralization information. The main phonological attributes are shown in Table 2.

Data field	Value
Diacritized	مُحَمَّدٌ
Phonemic (CARS)	*muhammadun*
Phonetic (IPA)	[muˈhɛ̆mmɛ̆dun]
X-SAMPA	mu"X\E_"mmE_"dun
Transliterated (Buckwalter)	muham∼adN

Table 2: Phonological attributes for مُحَمَّدٌ

3.2.3. Morphological and Orthographic Attributes

The morphological attributes include all features needed to obtain a given form out of its lemma. They are useful for morphological analysis, semantic analysis, lemmatization, decliticization, deaffixation, verb conjugation, and dictionary lookup. Operations such as decliticization, deaffixation, and tokenization (Carbonell et al., 2006) are easy to perform since clitics are given explicitly in their own fields (Enclitic, Proclitic, and Stem). The main morphological attributes are presented in Table 3.

Data Field	Value	Transcription
Full-form	ولكاتبكما	*walikātibíkumạ*
Lemma	كاتب	*kắtibun*
Stem	كاتب	*kắtib*
Proclitic	ول	*wali*
Enclitic	كما	*(i)kúmạ*
Root	ك-ت-ب	*k-t-b*

Table 3: Morphological attributes

Orthographic attributes are useful for orthographic disambiguation, which is necessary for word and entity recognition, TTS, morphological analysis, normalization, and dictionary lookup. These include orthographic variants, such as pausal and elided forms, as well as common typographical oddities. Here is an example of typical orthographic variants for the given name Alexandra: ألكسندرا, الكسندرا, ألكسندره, الكسندره, ألكسندرة, الكسندرة. As can be seen in these examples, ه and ة are sometimes interchangeable in names.

Orthographic variants also include allographs, for example, the use of ى (*alif maqsuura*) as an alternative for ي (*yaa*) in Egypt (Pinon, 2017), and the use of پ instead of ب for [p] in some regions.

3.3. Applications

3.3.1. Arabic Speech Technology

Due to the considerable ambiguity due to the graphemic underrepresentation of Arabic, even major IT players struggle to synthesize speech accurately. A survey (Halpern, 2020) revealed that it is not unusual for over 50%, and even 80%, of the words in a sentence to be mispronounced, especially cliticized words. In that survey, a pronunciation is considered erroneous when it includes mistakes such as incorrect case endings (e.g., pronouncing الكاتب as *lkắtibi* in an MSA context when it should be *lkắtibu),* omitted *shadda* (such as pronouncing عدد as *ɛádada* when it should be *ɛáddada* 'to enumerate'), or other pronunciation errors that can be unambiguously identified. In Table 4, pronunciation errors are marked by an asterisk.

Unvocalized	Vocalized	Google (13%)	iOS (31%)	Bing (25%)	CJKI
عدد	عَدَّدَ	*ɛádadu	*ɛádada	*ɛádada	ɛáddada
الكاتب	ٱلْكَاتِبُ	*lkấtibi	lkấtibu	lkấtibu	lkấtibu
ما	مَا	*mạ	mạ	mạ	mạ
الحكام	ٱلْحُكَّامَ	*lḥukkấmi	*lḥukkấmi	*lḥukkấmi	lḥukkấma

Table 4: Mispronunciations in composed text

3.3.2. Phonemic and Phonetic Transcription

ArabLEX addresses these shortcomings by serving as a comprehensive pronunciation dictionary to enhance the quality of both text-to-speech (TTS) and automatic speech recognition (ASR). It includes an NLP-oriented morpho-phonemic transcription called CARS (Halpern, 2009a), which accurately represents Arabic phonemes, while also encoding morphological information such as vowel neutralization. In addition, two phonetic transcriptions—SAMPA (Wells, 1997) and IPA (International Phonetic Association, 1999)—are provided to ensure accurate phonetic realizations.

Even fully vocalized Arabic does not always represent Arabic phonology correctly, let alone phonetics and prosody (Halpern, 2009a). In ArabLEX, the starting point for generating accurate phonemic and phonetic transcriptions is based on vetted vocalization (full *tashkiil*). Rigorously tested algorithms generate both phonemic (CARS) and phonetic (IPA) transcriptions that represent actual realizations with precision empirically verified to be near-perfect. The availability of phonetic transcriptions is particularly relevant for ASR systems, as phonetics can be utilized to improve them (Feng et al., 2023).

3.3.3. Automatic Speech Recognition

Indeed, ASR systems must recognize alternative pronunciations, including informal ones. For example, the MSA pronunciations of كاتبون 'writers' and أكتب 'I write' are *kạtibũna* and *ʾáktubu,* but the less formal variants *kạtibũn* and *ʾáktub* are very widespread. Such alternatives include pausal forms and final vowel elision. The former refers to sentence-final forms causing final vowels to be elided in Classical Arabic. At the same time, the latter is the elision of certain final vowels in both medial and final forms, common in spoken MSA and dialects. For example, رَجَعْتُ إِلَى ٱلْبَيتِ 'I returned home', pronounced *rajáɛtu ʾílạ_lbáyti,* in pausal form becomes *rajáɛtu ʾílạ_lbayt* and in spoken MSA becomes *rajáɛt ʾílạ_lbayt*. Note how the final *ti* and *tu* are truncated to *t*.

3.3.4. Named Entity Recognition

The DAN module of ArabLEX covers about 100,000 vocalized personal names and their 6.5 million romanized variants. DAN is widely deployed in both security and NLP processing tools for NER and MT. Similarly, the DAF and DAP modules comprise approximately 240,000 names for places and non-Arabic personal names. These modules account for about 450 million fully inflected and cliticized entries in ArabLEX (Halpern, 2009b).

3.3.5. Machine Translation

Although machine translation has dramatically improved translation quality, it has some shortcomings (Koehn, 2020). Some issues in Arabic are (1) the high orthographic ambiguity, (2) the morphological complexity (forms like ولكاتباتهما are difficult to analyze), (3) the recognition of named entities (often cliticized), and (4) the large number of word-forms for nouns and verbs. ArabLEX offers comprehensive coverage of inflected and cliticized forms and can be used to supplement existing corpora or as a pseudo-corpus for language model training. Additionally, the proper noun modules of ArabLEX, representing the most comprehensive collection of native and foreign proper nouns to our knowledge, are bilingual and romanized, serving as a bilingual dictionary.

4. Compilation Methodology

4.1. Sources and Reference Materials

The ArabLEX team, consisting of experienced lexicographers, professional editors, computational linguists, and software engineers, used various sources and reference materials to compile, validate, and proofread the data. This includes *The CJKI Arabic Learner's Dictionary* (The CJK Dictionary Institute, 2025b), *Frequency Dictionary of Arabic*, the *CJKI Arabic Verb Conjugator* (CAVE) (The CJK Dictionary Institute, 2011), the *Oxford Arabic Dictionary* and Wiktionary, as well as resources compiled by our institute and dozens of consultants, including dictionary companies in the Middle East. These resources are based on such corpora as the *Oxford Arabic Corpus*, the multi-million-word

MSA corpus by Mohammed Attia (Attia et al., 2011) and the 30-million-word Buckwalter-Parkinson MSA corpus (Buckwalter, 2011).

Proper nouns were drawn from the following resources created by the CJK Dictionary Institute: Database of Arabic Names, Database of Foreign Names in Arabic, Database of Arabic Place Names, which contains millions of personal and place names and their variants.

4.2. Quality Control

It can be argued that generating entries by rules and templates can result in non-existent or erroneous forms. The ArabLEX team conducted extensive research to ensure maximum accuracy and comprehensive coverage of all wordforms and their variants. Many programs were developed for data validation and proofreading to ensure accuracy and consistency, including those for automatic error detection and correction. The following summarizes the data validation process used by the ArabLEX team to refine the vocalization validation module (VBW_INTEG) and ensure accurate, fully vocalized Arabic and phonemic transcriptions for speech technology: (1) A program validates the correct vocalization of inflections, based on strictly defined rules such as *hamz*a rules, the presence of short vowels, and many more. (2) The program then attempts to rectify the errors it encounters autonomously. (3) Errors that the program cannot rectify are presented to proofreaders, who manually classify, analyze, and rectify them. (4) Based on the feedback of proofreaders, the validation rules are then either adjusted or the database of exceptions is expanded. (5) The process is then repeated.

This iterative process has been applied over the course of many years, resulting in a system with a comprehensive, reliable, and explainable set of rules and exceptions.

4.3. Inflection, Conjugation, Cliticization

Generating inflected forms involves many complex steps, including sanity checking and human proofreading. Nouns and adjectives are declined/inflected for feminine, dual, and plural forms. For example, for /bayotN/ 'house,' we derive /bayotaAni/, /buyuwtN/, and /buyuwtaAtN/.

The verb paradigms from CAVE are used to acquire the verb conjugations for each subject pronoun for each tense. CAVE has 180 categories and fully explicit, hand-vetted conjugations for each category. For example, for /kataba/ 'he wrote' we get /yakotubu/ (third person masculine singular imperfect), /Aukotubo/ (second person masculine singular imperative), etc. To encliticize, the correct enclitic template is selected based on the ending of the inflected form. For example, the noun /|xirapu/ 'the hereafter' ends in /pu/, so the template in Table 5 is selected. Enclitics are then added to correspond to each case and subject pronoun. For /|xirapu/, we generate such forms as /|xiratiy/, /|xiratuka/ and /|xiratuki/. To procliticize, the appropriate proclitics are selected from the template. For example, for /bayotN/ 'house', the enclitic is /-N/ (tanwiin), so we refer to the appropriate row (row 2) in Table 6 and generate />abayotN/, /wabayotN/, etc.

The clitics are not merely blindly concatenated to the base form. There are over 2,000 valid orthographic, grammatical, and semantic combinations of clitics, as defined by ArabLEX's human-vetted constraint-defining tables (shown below), and several thousand that are invalid.

Per	Case	Enclitic	Rule
000	NOM	u	
1SC	NOM	iy	-p → -t
2SM	NOM	uka	-p → -t
2SF	NOM	uki	-p → -t

Table 5: Template for nouns that end in /p/ ة

Proclitic	Enclitic	Gen	Num
0, >a, wa, fa, >awa, >afa, Aalo, ...	a, u	M	S
0, >a, wa, fa, >awa, >afa	N, FA, FY	M	S
0, >a, wa, fa, >awa, >afa	uhaA, uhu, uhumaA, uhumo, uhun∼a, uka, uki, ukumaA, ...	M	S

Table 6: Possible combinations of clitics

5. Comparison with Other Resources

5.1. Comparing Resources

Previous efforts to compile extensive Arabic lexicographical or lexical databases have yielded datasets containing around 200,000 unique lemmata. These datasets tend to lack a diverse set of attributes. By contrast, detailed datasets typically contain around 30,000 unique lemma entries, e.g., the CALIMA dataset for Egyptian Arabic (Alshargi et al., 2019).

ArabLEX, on the other hand, covers a combined 375,335 unique lemmata, including a large number of named entities, while exceeding the level of detail of its counterparts, especially by offering phonetic (IPA, XSAMPA) and phonemic (CARS) transcriptions and fully diacritized Arabic.

Another advantage of ArabLEX is the total number of entries accessible for explicit analysis; that is, entries that are pre-generated as opposed to on-the-fly generation. For example, the CALIMA dataset contains approximately 48 million entries that can be obtained when all supported wordforms are exhaustively generated (AlShuhayeb et al., 2023). By contrast, ArabLEX consists of 570 million pre-compiled entries, immediately accessible for use and analysis.

As Camel Morph is a very recent resource that is in many ways comparable to ArabLEX, we have devoted the following section to a detailed comparison of these two resources.

5.2. Comparison with Camel Morph

ArabLEX is more refined than other morphological engines. To illustrate this, we compared the principal features of ArabLEX against Camel Morph (Khairallah et al., 2024), an advanced, comprehensive engine comparable in coverage and features to ArabLEX. The comparison below is between ArabLEX v1.2 and Camel Morph 2024. It demonstrates that although Camel Morph is a full-fledged, large-scale resource with many features, ArabLEX has various features that surpass it in both quantity and quality, such as proper nouns and more refined POS categories.

5.2.1. Coverage

Since the structure and format of ArabLEX and Camel Morph are fundamentally different, the number of entries is calculated differently and is not strictly comparable. ArabLEX v1.1 covered approximately 530 million entries as compared to Camel Morph's 535 million, indicating they have almost identical coverage. ArabLEX v1.2, however, includes about 14,000 new canonical headwords, so that the total number of entries attains approximately 570 million.

The total number of all possible unique analyses in Camel Morph is over 1.4 billion. The total number of unique field values in ArabLEX is about 15 billion. As for proper noun lemmata, the difference in coverage is dramatic (see below).

5.2.2. POS coverage

For the principal content words (verbs, nouns, adjectives), both resources have comparable coverage, approximately as shown below. The numbers include all inflected and variant forms.

The number of function words is minuscule compared to content words. ArabLEX v1.2 includes full coverage of such word classes as prepositions and conjunctions, and their inflections.

POS	ArabLEX v1.1	ArabLEX v1.2	Camel Morph
Nouns	14,293	22,911	19,965
Adjectives	6,228	9,893	7,205
Verbs	9,509	10,948	9,333
Total	30,030	43,752	36,503

Table 7: Content words

For proper nouns, the differences are dramatic. Camel Morph includes some 69,558 lemmata for proper nouns of unknown type. ArabLEX covers 345,000 lemmata and a total of 451 million inflections (cliticized and inflected forms), as shown below.

Type	Subtype	Lemmata	Inflections
Anthroponyms	Arab	100,312	218,215,875
	non-Arab	223,367	226,784,907
Toponyms	Arab	14,804	4,424,174
	non-Arab	6,822	2,031,027

Table 8: Proper nouns in ArabLEX

Breakdown by type data for Camel Morph is not available, but the total number of proper noun inflections is around 57 million, as opposed to 451 million in ArabLEX.

5.2.3. Missing and Rare Lemmata

We conducted a mutual comparison of missing lemmata in both resources and found thousands of rare entries in Camel Morph. Because of its complex structure, it is difficult to calculate exact figures (the many misspelled and non-existing heawords in Camel Morph skew the count). We checked missing lemmata against the set of standard dictionaries used in compiling ArabLEX and examined their web frequencies. We estimate that ArabLEX contains approximately 11,000 lemmata not attested in Camel Morph, including thousands of common words such as حَصَّلَ 'obtain', وَفِقَ 'agree', and كَيَّفَ 'adopt, adjust', all with high web frequencies.

Camel Morph contains approximately 17,000 lemmata missing in ArabLEX v1.1. Some are archaic, erroneous, or rare words not attested in this set of dictionaries, such as أَمْحَكَ, خَاوَى, تَغَرَّقَ. This may reflect tokenization errors or other artifacts introduced during corpus processing. Most of these missing lemmata have been incorporated into ArabLEX v1.2 after undergoing manual vetting to ensure their validity.

Arabic	CAPHI	CARS	Proxy CARS	IPA
طَالِبٌ	t. aa l l b u n	ṭā́libun	Taa/libun	ˈtˤɑːlibun
اَلْيَابَانُ	2 a l y aa b aa n u	ʾalyaa̲bā́nu	'alya_baa/nu	ʔaljaˈbɑːnu
كِيلُوغْرَامٌ	k ii l uu gh r aa m u n	kīlūghrā́mun	ki_lu_ghraa/mun	kiluɣˈrˤɑːmun
قَصِيدَةٌ	q a s. ii d a t u n	qaṣī́datun	qaSii/datun	qɑˈsˤiːdatun
صُبْحٌ	s. u b 7 u n	ṣúbḥun	Su/bHun	ˈsˤʊbħun

Table 9: CARS vs. CAPHI

5.2.4. POS Classification

A notable limitation of Camel Morph's part-of-speech classification scheme is the lack of explicit codes to indicate verbal nouns, active participles, passive participles, and *nisba* (adjectival nouns). Since in Camel Morph, they are all treated as nouns or adjectives, there is no way to extract them or make direct queries on them.

In ArabLEX, these forms are treated explicitly as SUBPOS categories of verbs, adjectives, or nouns. For example, as a verb, the active participle كَاتِب *kā́tibun* denotes an ongoing process, 'I am writing' or 'I write.' As an adjective, it functions like the English present participle to modify a noun, as in رَجُلٌ نَائِمٌ *rájulun nā́ʾimun* 'a sleeping man.' As a noun, it designates the agent, as in كَاتِبٌ *kā́tibun* 'writer, author'. By contrast, Camel Morph does not encode such information.

As for verb transitivity, ArabLEX has explicit SUBPOS codes for transitive, intransitive, and ambitransitive verbs so that one can access these subcategories directly. The Camel Morph analyzer does not display transitivity codes.

5.2.5. English glosses

ArabLEX consists of four modules. The DAG (Arabic general vocabulary) module does not yet include English glosses, because the goal of ArabLEX is to provide morphological, phonetic, and grammatical information, rather than act as a translation dictionary. However, in the three modules for proper nouns, over 330,000 English or romanized glosses and/or romanized variants are provided.

5.2.6. Speech technology

ArabLEX places special focus on speech technology by providing various features that strongly support TTS and ASR applications. For all 570 million entries, it provides accurate vocalization, **phonetic** transcriptions (IPA and SAMPA), and CARS **phonemic** transcriptions (Halpern, 2009a) that indicate vowel neutralization, word stress, and velarization explicitly. ArabLEX has significantly contributed to Amazon's advanced speech technology for Alexa to achieve error reduction for both speech synthesis and voice recognition.

5.2.7. CARS vs. CAPHI

Camel Morph includes a transcription feature called CAPHI (Habash et al., 2018), a practical phonemic transcription (often mislabeled "phonetic") designed for both MSA and dialects. ArabLEX offers IPA, the international standard, and CARS, a robust *phonemic-enhanced* MSA transcription system that optionally adds important phonetic/allophonic realizations such as velarization.

CARS and CAPHI differ mainly in the prosodic and morpho-phonemic detail they encode. CARS explicitly marks word stress (acute accent), neutralization (macron below), and velarization ([ɑ]), going beyond simple phoneme mapping. CARS also offers liaison marking, optional syllable annotation, easy-to-type proxy symbols, and conversion tools, making it flexible for both human (pedagogy) and machine use (NLP). CAPHI, by contrast, provides a clear one-to-one sound mapping across dialects but does not mark these features, as can be seen in Table 9.

As can be seen in Table 9, though CARS is basically phonemic, it can (optionally) encode phonetic/allophonic realizations, such as in *ṭā́libun*, and neutralized vowels as in *ʾalya̲bā́nu*, where *a̲* is a neutralized long /a/ and *ā́* is a velarized stressed long a. Such rich phonemic and phonetic representations are not available in CAPHI.

5.2.8. Availability

While Camel Morph is available as open-source software through github, the ArabLEX core module(s) or subsets are available for free of charge for academic research. For commercial development, ArabLEX is available for licensing or through ELRA.

6. Future Work

After defining the notion of full-form lexicon, we have presented an Arabic full-form lexicon resource called ArabLEX, and provided information on its structure and applications. We have compared it with Camel Morph, a large-scale morphological generator.

The development of ArabLEX is ongoing. Expansion of ArabLEX will continue by adding new entries and data fields, including technical terms, named entities, more phonological attributes, orthographic variants, alternative pronunciations, and additional word classes. Especially noteworthy are new headwords that consist of multiword expressions (Halpern, 2019) (inflections or conjugations consisting of space-separated components), such as periphrastic elatives (أَكْثَرُ إِيلَامٍ 'more painful'), negative elatives (with أَقَلُّ or أَخَفُّ), inflected numerical expressions, phrasal verbs, compound tenses, verb negation, and more.

The addition of proclitics, enclitics, and inflections leads to ArabLEX exceeding about 570 million records. Eventually, ArabLEX is expected to reach about one billion records. We hope that ArabLEX will prove helpful to the community of Arabic speakers and that it will enhance the level of support available for the Arabic language and, therefore, will enhance its chances of survival and proliferation in years to come.

7. Ethics Statement

This work presents a large-scale lexical resource for Arabic that aims to advance natural language processing and computational linguistics research for a language community that language technologies have historically underserved. We recognize both the potential benefits and risks associated with developing and disseminating such resources, and address the following ethical considerations.

7.1. Linguistic Equity and Access

This resource helps bridge the technological gap between Arabic and better-resourced languages by providing comprehensive morphological and phonological annotations at scale. ArabLEX is available for both commercial distribution as well as for supporting academic research through a free licensing program for qualifying academic projects. We believe this model balances sustainability with broad accessibility for the research community.

7.2. Data Sources and Permissions

The lexical data in this resource is based on or uses for reference both proprietary sources, such as the CJKI Arabic Learner's Dictionary, commercial resources, such as the Oxford Arabic Dictionary, and publicly available resources and websites, such as Wiktionary and Almaany. All proprietary materials have been used in accordance with licensing terms and permissions. We acknowledge the contributions of the lexicographic community and the volunteer contributors to open resources that have made this work possible.

7.3. Representation and Scope

This resource primarily covers Modern Standard Arabic (MSA), the variety most widely used in formal contexts across the Arabic-speaking world. While our complete resource includes coverage of major dialects (The CJK Dictionary Institute, 2025a), the present work focuses on MSA. We acknowledge that this focus may not fully represent the linguistic diversity of Arabic speakers and that dialectal varieties spoken by millions remain less represented in computational resources. Future work should prioritize expanding coverage of these underrepresented varieties.

7.4. Potential Applications and Dual-Use Considerations

Lexical resources enable a wide range of beneficial applications, including machine translation, morphological analysis, named entity recognition, language learning tools, accessibility technologies for speakers with disabilities, and educational resources. However, we recognize that comprehensive linguistic resources can also be used in applications that raise ethical concerns, such as surveillance systems, automated content filtering that may enable censorship, or profiling tools. While we cannot control all downstream uses of commercial resources, we encourage users to consider the societal implications of their applications and to develop technologies that respect user privacy and human rights.

7.5. Licensing Model

By adopting a hybrid model of commercial licensing combined with academic licensing, we aim to ensure the long-term maintenance and continued development of this resource. Regular updates and error corrections are essential for maintaining data quality, and sustainable funding enables ongoing community engagement and responsiveness to user needs.

7.6. Broader Impact

We hope this resource will empower researchers, educators, and developers to create more sophisti-

cated and accurate Arabic language technologies, ultimately benefiting the hundreds of millions of Arabic speakers worldwide. We remain committed to engaging with the Arabic-speaking linguistic and technical communities to ensure that this resource serves their needs responsibly and effectively.

References

A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA*, pages 11–16. Association for Computational Linguistics.

F. Alshargi, S. Dibas, S. Alkhereyf, R. Faraj, B. Abdulkareem, S. Yagi, O. Kacha, N. Habash, and O. Rambow. 2019. Morphologically Annotated Corpora for Seven Arabic Dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147.

H. AlShuhayeb, B. Minaei-Bidgoli, M. E. Shenassa, and S. Hossayni. 2023. Noor-Ghateh: A Benchmark Dataset for Evaluating Arabic Word Segmenters in Hadith Domain. ArXiv preprint.

M. Attia, P. Pecina, L. Tounsi, A. Toral, and J. van Genabith. 2011. A lexical database for modern standard Arabic interoperable with a finite state morphological transducer. In *Systems and Frameworks for Computational Morphology. Second International Workshop, SFCM 2011, Zurich*, volume 100 of *Communications in Computer and Information Science*. Springer.

M. Boudchiche, A. Mazroui, M. Ould Abdallahi Ould Bebah, A. Lakhouaja, and A. 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.

T. Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. *2002*, LDC Catalog No.: LDC2002L49.

T. Buckwalter. 2011. *A Frequency Dictionary of Arabic. Core Vocabulary for Learners*. Routledge.

J. Carbonell, S. Klein, D. Miller, M. Steinbaum, T. Grassiany, and J. Frei. 2006. Context-Based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA*, pages 19–28. Association for Machine Translation in the Americas.

R. Cotterell and H. Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

J. Dichy. 2017. The Analytics of Writing, Exemplified by Arabic, the Youngest of the Semitic Scripts. In *Approaches to the History and Dialectology of Arabic, in Honour of Pierre Larcher*, pages 29–56, Leiden, Boston. Brill.

S. Feng, M. Tu, R. Xia, C. Huang, and Y. Wang. 2023. Language-universal phonetic encoder for low-resource speech recognition. arXiv. ArXiv preprint.

N. Gadelli. 2015. *The morphological integration of loanwords into Modern Standard Arabic: Towards a morphological categorization of loanwords*. Ph.D. thesis, Lund University, Sweden.

N. Habash, F. Eryani, S. Khalifa, O. Rambow, D. Abdulrahim, A. Erdmann, R. Faraj, W. Zaghouani, H. Bouamor, N. Zalmout, S. Hassan, F. Al-Shargi, S. Alkhereyf, B. Abdulkareem, R. Eskander, M. Salameh, and H. Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

N. Habash, O. Rambow, and R. Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS 150 tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo*.

J. Halpern. 2009a. CJKI Arabic Romanization System (CARS). In *Romanization of Arabic Names: Proceedings of the International Symposium on Arabic Transliteration Standard: Challenges and Solutions, Abu Dhabi, UAE, 15–16 December 2009*. Ministry of Culture, Youth and Community Development.

J. Halpern. 2009b. Lexicon-Driven Approach to the Recognition of Arabic Named Entities. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo*.

J. Halpern. 2009c. Word stress and vowel neutralization in modern standard Arabic. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April*. The MEDAR Consortium.

J. Halpern. 2019. Lexicographic Criteria for Selecting Multiword Units for MT Lexicons.

J. Halpern. 2020. Enhancing Arabic Speech Technology with Comprehensive Arabic Training Lexicon.

Y. Haralambous. 2021. Breaking Arabic: the creative inventiveness of Uyghur script reforms. *Design Regression*.

Y. Haralambous. 2024. *A Course in Natural Language Processing*. Springer.

E. Haugen. 1972. Schizoglossia and the Linguistic Norm. In *Studies by Einar Haugen*, pages 441–445. De Gruyter Mouton, Berlin, Boston.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Ch. Khairallah, S. Khalifa, R. Marzouk, M. Nassar, and N. Habash. 2024. Camel morph MSA: A large-scale open-source morphological analyzer for Modern Standard Arabic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2683–2691, Torino, Italia. ELRA and ICCL.

P. Koehn. 2020. *Neural Machine Translation*. Cambridge University Press, Cambridge, UK.

X. Lu, P. West, R. Zellers, R. Le Bras, C. Bhagavatula, and Y. Choi. 2021. Neurologic decoding: (un)supervised neural text generation with predicate logic constraints. ArXiv preprint.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt*, pages 102–109.

G. Mejdell. 2014. Luġat al-ʾumm and al-luġa al-ʾumm - the 'mother tongue' in the Arabic context. In *Arabic and Semitic Linguistics Contextualized*, pages 214–226. Harrassovitz.

D. Meletis and Ch. Dürscheid. 2022. *Writing Systems and Their Use. An Overview of Graphonlinguistics*, volume 369 of *Trends in Linguistics*. de Gruyter.

J.R. Osborn. 2017. *Letters of Light. Arabic Script in Calligraphy, Print, and Digital Design*. Harvard University Press.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, May*. European Language Resources Association (ELRA).

C. Pinon. 2017. Intégrer les variations dans l'enseignement de l'arabe langue étrangère: enjeux et méthodes. In *Arabe standard et variations régionales. Quelle(s) politique(s) linguistique(s)? Quelle(s) didactiques)?* Éditions des archives contemporaines.

K. C. Ryding. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge, UK.

M. S. S. Sawalha. 2011. *Open-source resources and standards for Arabic word structure analysis: Fine-grained morphological analysis of Arabic text corpora*. Ph.D. thesis, The University of Leeds School of Computing.

R. Sennrich, B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

O. Smrž. 2007. ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic*, pages 1–8. Association for Computational Linguistics.

A. Soudi and A. Eisele. 2004. Generating an Arabic Full-form Lexicon for Bidirectional Morphology Lookup. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal*. European Language Resources Association (ELRA).

M. Tahon, R. Qader, G. Lecorvé, and D. Lolive. 2016. *Improving TTS with corpus-specific pronunciation adaptation*. Interspeech, San Francisco, CA.

D. Taji, S. Khalifa, O. Obeid, F. Eryani, and N. Habash. 2018. An Arabic Morphological Analyzer and Generator with Copious Features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology, Brussels*, pages 140–150. Association for Computational Linguistics.

The CJK Dictionary Institute. 2011. The CJKI Arabic Verb Conjugator.

The CJK Dictionary Institute. 2025a. DiaLEX: Arabic Dialects Full-form Lexicon.

The CJK Dictionary Institute. 2025b. The CJKI Arabic Learner's Dictionary.

J. C. Wells. 1997. SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter. Part IV, section B, Berlin and New York.

Q. Zhu, X. Hu, P. Ji, W. Wu, and K. Tu. 2024. Unsupervised morphological tree tokenizer. ArXiv preprint.

8. Language Resource References

The CJK Dictionary Institute. 2019a. *ArabLEX: Database of Arab Names (DAN)*. distributed via ELRA: ELRA-Id ELRA-M0107, Lexicon, 1.0, ISLRN 773-974-582-139-4.

The CJK Dictionary Institute. 2019b. *ArabLEX: Database of Arabic General Vocabulary (DAG)*. distributed via ELRA: ELRA-Id ELRA-L0131, Lexicon, 1.0, ISLRN 879-334-992-724-8.

The CJK Dictionary Institute. 2019c. *ArabLEX: Database of Arabic Place Names (DAP)*. distributed via ELRA: ELRA-Id ELRA-M0105, Lexicon, 1.0, ISLRN 161-842-321-771-2.

The CJK Dictionary Institute. 2019d. *ArabLEX: Database of Foreign Names in Arabic (DAF)*. distributed via ELRA: ELRA-Id ELRA-M0106, Lexicon, 1.0, ISLRN 943-592-129-040-2.