日中韓辭典研究所
THE CJK DICTIONARY INSTITUTE, INC.

ArabLEX vs. Camel Morph: Enabling Real-World Arabic NLP

Comparing the leading lexical resources in Arabic NLP

by **Jack Halpern**

November 17, 2025

0. Introduction

The **Arabic Full Form-Form Lexicon** (ArabLEX) is a full-form lexicon (FFL) designed for industrial-strength Arabic NLP and speech technology. **Camel Morph MSA** (2024) is currently an open-source MSA morphological analyzer/generator. Both of these resources can be used for academic research. This report describes how ArabLEX can be used as a highly effective computational lexicon for NLP applications, offers a detailed comparison of these two resources, and highlights their major features.

Note that is a comparative study report, not an academic paper. Thus various statements are made without citing evidence and without citations.

1. About CJKI

Founded in 1993 in Japan, CJKI builds large-scale lexical resources for Chinese, Japanese, Korean, Arabic, and other languages. It specializes in large-scale lexical databases that power machine translation, speech technology, as well as pedagogical applications (dictionaries and apps), and compliance (e.g., AML/KYC name matching).

2. Technical Background

Unlike conventional dictionaries that record only canonical lemmata, a **full-form lexicon** (FFL) aims to exhaustively cover declined, conjugated and cliticized (“inflected” for short) wordforms of a language. A comprehensive FFL plays a central role in resolving the high level of morphological, orthographic, and phonological ambiguities inherent in the Arabic writing system.

Unlike **morphological generators** such as Camel Morph, which dynamically generate forms based on rules and supporting databases, an FFL aims to provide the user with all pre-computed wordforms stored in a database. This provides faster lookup, a clear format and structure, greater SQL compatibility, and easy integration into large language models.

3. Overview of ArabLEX

To address ambiguities inherent in unvocalized written Arabic, ArabLEX offers a full-form lexicon of Modern Standard Arabic (MSA) that aims for complete morphological and phonological coverage. Now exceeding 570 million entries (about 15 billion field values), it is the largest Arabic lexical resource available. ArabLEX also serves as a foundational framework of **DiaLEX**, a database of some of the major Arabic dialects totaling about 150 million entries (soon to double in size).

ArabLEX consists of four primary modules:

DAG – Arabic General Vocabulary (83 million entries)
DAN – Arabic Names (218 million entries)
DAF – Arabic Foreign Names (226 million entries)
DAP – Arabic Place Names (6 million entries)

These modules provide up to 25 linguistic fields per entry, covering grammatical, phonological, morphological, and orthographic attributes. Stored in UTF-8 TSV format, ArabLEX is simple to query, to integrate into SQL systems, to use as a database for morphological engines, and to integrate into LLMs. The data fields depend on the module. The following table illustrates ArabLEX's explicit representation of proclitics and enclitics, a key feature for accurate tokenization and morphological analysis.

Data Field	Value	CARS 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 1: Morphological attributes

The high-precision morphophonemic data and three transcription systems (CARS, IPA, SAMPA) support both ASR and TTS, ensuring accurate recognition and natural speech output. At the same time, the massive bilingual inflectional coverage of inflected proper nouns enhances named-entity recognition and machine translation while effectively resolving orthographic and morphological ambiguity. To that end, subsets of ArabLEX, especially DAN (Arab names), are currently used by software developers of NLP tools, especially security applications related to money laundering and terror watchlists.

4. Overview of Camel Morph

Several tools have been developed for morphological analysis and generation for Arabic NLP, including AlKhalil, MADA, BAMA, PATB, FARASA, MADAMIRA, Elixir_FM and CALIMA Star. The most recent, Camel Morph MSA, is an open-source database for morphological analysis and generation developed at NYU Abu Dhabi's CAMeL Lab. The underlying database links prefixes, stems, and suffixes through compatibility rules, and is accessed by the Camel Tools API. It also incorporates detailed morphological specifications, orthophonological rewrite rules, and the CAPHI transcription scheme.

The database covers about 105,000 lemmata, 1.45 billion analyses, and approximately 535 million unique diacritizations, providing the most comprehensive open-source resource for Modern Standard Arabic. While it does not provide full-form entries like ArabLEX, Camel Morph's rule-based structure makes it a versatile platform for linguistic research, annotation, and the study of Arabic morphology.

5. ArabLEX vs. Camel Morph

ArabLEX and **Camel Morph** are both large-scale Arabic morphological resources, but they differ in design philosophy, structure, and application. The comparison below is based on ArabLEX v1.2 and Camel Morph MSA (2024).

5.1 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 included about 530 million entries, whereas Camel Morph covers 535 million entries, effectively equivalent. Camel Morph includes about 110,000 lemmata, whereas ArabLEX v1.2 contains approximately 390,000 lemmata (due to the large number of named entities), bringing the total to 570 million entries.

5.2 POS Coverage

For nouns, adjectives, and verbs, the number of canonical forms for both resources are roughly equivalent.

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 2: Content words

For proper nouns, the differences are dramatic. Camel Morph covers about 69,558 entries, while ArabLEX includes 345,000 lemmata with 451 million inflections/clitics forms, as 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 3: Proper nouns in ArabLEX

By contrast, Camel Morph's total inflections for proper nouns is around 57 million, eight times less.

5.3 Missing and Rare Lemmata

A bidirectional audit of missing lemmata revealed that ArabLEX contains numerous high-frequency items that are missing in Camel Morph,. The Google hits in the two tables below was performed on *vocalized* Arabic using quoted keywords via the Google API.

Arabic	CARS	Google hits
أَنَسَ	*ʾánasa*	31,100
حَصَّلَ	*ḥáṣṣala*	7,270
عَلَّ	*ɛálla*	165,000
فَرَجَ	*fáraja*	136,000
كَيَّفَ	*káyyafa*	2,480
وَفِقَ	*wáfiqa*	1,070

Table 4: Lemmata only in ArabLEX

These omissions might result from tokenization errors or rule-generation limits. Conversely, Camel Morph has some obscure or erroneous forms that are not supported by standard dictionaries, as below.

Arabic	CARS	Google hits
أَرْتَقَةٌ	*ʾartáqatun*	0
أَمْحَكَ	*ʾámḥaka*	54
إِكْتُرُونِيٌّ	*ʾiktṟuníyyun*	0
إِنْقِلِيزِيٌّ	*ʾinqlḭzíyyun*	0
تَغَرَّقَ	*taghárraqa*	5
خَاوَى	*khā́wạ*	7

Table 5: Lemmata only in Camel Morph

The "Google hits" column shows that Camel Morph is missing many common words while it contains numerous (thousands of) rare or non-existing words. Most of these legitimate missing lemmata have since been incorporated into ArabLEX v1.2 after rigorous manual vetting so to ensure that it covers real-world usage based on frequency statistics and grammatical validity.

5.4 POS Classification

A notable limitation of Camel Morph lies in part-of-speech (POS) granularity, which lacks explicit POS codes to indicate verbal nouns, active participles, passive participles, and *nisba* (adjectival nouns). Since it lumps them together under the noun and adjective categories, it obscures their syntactic/morphological distinctions. For instance: as a verb, كَاتِب (*kā́tibun*) denotes an ongoing process (“is writing”); as an adjective, it modifies a noun (“a writing man”); and as a noun, it denotes the agent (“writer, author”).

In ArabLEX, by contrast, these distinctions are represented explicitly by SUBPOS tags, allowing precise filtering and retrieval. Similarly, ArabLEX encodes verb transitivity with sub-tags for transitive, intransitive, and ambitransitive verbs—categories not given by the Camel Morph analyzer. This enhanced typology enables ArabLEX users to query at a deeper morphological depth to determine morphological and syntactic behavior.

5.5 English Glosses

Although the DAG (General Vocabulary) module focuses on morphological rather than semantic information, the other three modules—DAN, DAF, and DAP—collectively supply over 330,000 English in addition to numerous romanized variants. These cover person and place names, as well as non-Arab proper nouns, thereby supporting bilingual NLP tasks such as machine translation and named-entity recognition.

5.6 Speech Technology

ArabLEX has been designed for phonological precision: it provides full vocalization, IPA, SAMPA, and CARS transcriptions, and Buckwalter (BW) transliteration for all 570 million entries. These explicitly mark vowel neutralization, word stress, and velarization, enabling accurate speech synthesis and recognition. These features are crucial, as for example in Amazon's Alexa, where integration of ArabLEX data resulted in significant error reduction in both TTS and ASR. By contrast, Camel Morph provides only partial phonemic data through its CAPHI transcription system (see next section) and is, therefore, of limited applicability to speech technology.

In fact, romanization errors are widespread among the major platforms (see Table 6), which could be avoided by utilizing ArabLEX's pronunciation dictionaries.

Unvocalized	Vocalized	Google (13%)	iOS (31%)	Bing (25%)	CJKI
عدد	عَدَّدَ	*εά-dadu	*εά-dada	*εά-dada	εά-ddada
الكاتب	ٱلْكَاتِبُ	*lkā́tibi	lkā́tibu	lkā́tibu	lkā́tibu
ما	مَا	*mạ	*mạ	*mạ	mạ
الحكام	ٱلْحُكَّامَ	*lḥukkā́ mi	*lḥukkā́ mi	*lḥukkā́ mi	lḥukkā́ ma

Table 6: Mispronunciations in composed text

(* pronunciation errors are marked by an asterisk)

5.7 CARS vs. CAPHI

Camel Morph's **CAPHI** system provides a unified phonemic representation across MSA and dialects, but it omits prosodic and morpho-phonemic features. ArabLEX's **CARS** system (CJKI Arabic Romanization System) goes further, combining phonemic accuracy with optional phonetic and prosodic detail. Key distinctions include:

- CARS explicitly marks stress, vowel neutralization, and velarization ([α]); CAPHI does not.
- CARS supports liaison marking, syllable segmentation, and proxy notation for easy typing.
- CARS is convertible to IPA and SAMPA, facilitating both pedagogical and computational use.
- CARS encodes *some* phonetic/allophonic realizations, such as in *ṭā́libun*, and neutralized vowels as in *ʾalyạbā́nu*, where ạ is a neutralized long /a/ and á is a velarized stressed long /a/. Such rich representations are not available in CAPHI.

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

Table 7: CARS vs. CAPHI

CARS thus offers significant advantages in enabling accurate grapheme-to-phoneme alignment.

5.8 Accessibility

Distribution format: ArabLEX is distributed as (1) simple UTF-8 TSV text files and (2) a MySQL database with a user interface. In contrast, Camel Morph is released as a set of internally structured database files.

Querying: For generation, analysis and querying, ArabLEX can be utilized flexibly through (1) a user interface, (2) SQL queries, or (3) a Python API. For Camel Morph, these tasks are realized through CAMeL Tools, which also provides a Python package, but does not provide a user interface.

Direct access: Since its format is not transparent, to access Camel Morph's internal text data directly (without CAMeL Tools) is not trivial and requires significant effort, while direct access to ArabLEX's text data is straightforward because of its simple structure (flat TSV text files).

Versatile retrieval: Unlike Camel Morph, the ArabLEX MySQL system allows versatile retrieval based on arbitrary search criteria, and is not limited to generation and analysis. On the other hand, Camel Morph's search parameters are limited by the constraints of the feature dictionary. As an example, a MySQL query in ArabLEX can request retrieval of "all transitive verbs beginning with أ of the third person singular or plural with ease," which cannot be done in a Camel Morph .

5.9 LLM Integration

With ArabLEX's simple TSV format, each line represents a complete set of wordforms based on a single lemma. This self-contained structure allows a simpler process of direct integration into LLM pipelines without reconstructing morphological combinations at runtime. By contrast, Camel Morph's rule-based design of complex file structure requires compilation and large-scale processing that is less efficient and less flexible than a TSV format. ArabLEX's compact, ready-to-use format is thus far better suited for scalable neural and retrieval-augmented applications.

5.10 Performance

ArabLEX stores over 570 million precomputed wordforms with detailed grammatical, phonological, and orthographic data. This structure enables efficient real-time processing, such as instant lookup, SQL integration, and seamless compatibility with LLMs.In contrast, Camel Morph MSA employs a rule-based morphological framework that links prefixes, stems, and suffixes through compatibility rules, performing analysis and generation at runtime via the interactive CAMeL Tools interface. While highly effective for linguistic analysis and experimentation, Camel Morph operates more slowly than lexicon-based systems such as SAMA or CALIMA. Consequently, when accessed through an API, its runtime performance is slower than that of morphological engines built on an integrated lexicon such as

ArabLEX. Its architecture is therefore less suited to industrial-grade, high-throughput systems that require fast, full-form lookup.

5.11 Availability

Camel Morph is open-source on GitHub, facilitating experimentation but with no commercial-grade support. ArabLEX, by contrast, is available free for academic research (core modules or subsets) and commercially licensed through ELRA or directly from CJKI, with commercial-grade support. Its dual-licensing model ensures both scholarly access and enterprise-level reliability—balancing openness with the quality assurance required for mission-critical NLP and speech applications.

6. Executive Summary

This report compares ArabLEX, a full-form lexicon (FFL) for industrial-scale Arabic NLP and speech technology, with Camel Morph, a rule-based morphological analyzer for MSA. Both address Arabic's complex morphology, but differ sharply in architecture and application.

ArabLEX stores over 570 million precomputed wordforms, enabling instant lookup, efficient processing in real-time and seamless integration into SQLs and LLMs. In contrast, Camel Morph is a rule-based framework that links affixes with stems through compatibility rules, performing analysis and generation via an interactive user interface. While highly effective for linguistic exploration, it has a slow runtime performance when accessed through an API. Therefore, its architecture is less efficient for high-throughput systems that depend on full-form lookup.

Both resources provide similar coverage, but ArabLEX offers vastly richer treatment of named entities. It also includes numerous lemmata missing from Camel Morph, which contains numerous rare or erroneous forms. ArabLEX also offers a more fine-grained POS tagging, marking participles, verbal nouns, and transitivity explicitly, and delivers three highly accurate phonemic and phonetic transcriptions at a phonological depth unavailable in Camel Morph.

ArabLEX's TSV format allows direct ingestion into neural and retrieval-based systems, while Camel Morph's multi-table rule-driven design is designed for runtime generation and heavier computation. In essence, Camel Morph excels as a linguistic laboratory, while ArabLEX delivers an industrial-strength lexicon optimized for real-world NLP easily integrated in LLMs and speech technology applications.

Reference Links

1. *ArabLEX technical specifications:* https://www.cjk.org/wp-content/uploads/Arablex_specs.pdf
2. *ArabLEX webpage:* https://www.cjk.org/data/arabic/nlp/arablex-arabic-full-form-lexicon/
3. *ArabLEX white paper summary:* https://www.cjk.org/wp-content/uploads/ArabLEX_summary_.pdf
4. *Camel Morph paper:* https://aclanthology.org/2024.lrec-main.240.pdf
5. *CAPHI summary:* https://camel-guidelines.readthedocs.io/en/latest/phonology/
6. *CARS paper:* https://www.cjki.org/arabic/cars/cars_paper.pdf

APPENDIX

DiaLEX: ARABIC DIALECTS FULL-FORM LEXICON

DiaLEX is the most comprehensive computational lexicon ever created for Arabic dialects. Designed for NLP applications like MT, NER and morphological analysis, it is ideally suited for training speech technology models.

No.	*Name*	*Full*
01	**EA_LEX**	Egyptian Arabic Full-Form Lexicon
02	**HA_LEX**	Hijazi Arabic Full-Form Lexicon
03	**LA_LEX**	Emirati Arabic Full-Form Lexicon
04	**PA_LEX**	Palestinian Arabic Full-Form Lexicon
05	**SA_LEX**	Syrian Arabic Full-Form Lexicon
06	**UA_LEX**	Emirati Arabic Full-Form Lexicon

Table 1: Dialects currently available or in progress

Coverage

DiaLEX covers some of the major Arabic dialects, including Egyptian, Saudi Arabian (Hijazi), and Emirati, as well as Palestinian, Syrian, and Lebanese - the latter three of which are under development.

Dialect	Lenmata	Entries
Egyptian	33,000	93 million
Hijazi	30,000	25 million
Emirati	29,000	37 million

Rich in morphological, grammatical, phonological, and orthographic attributes, DiaLEX maps all unvocalized forms to their vocalized counterparts and to the lemma, and provides transcriptions and transliterations. The number of entries in the table does not include *proclitics* (cliticized prefixes), which total several hundred million for the three dialects shown.

Distinctive Features

- Extremely comprehensive full form entries
- Rich in morphological attributes: all inflected, cliticized, and negated forms
- Numerous orthographic variants
- Includes high frequency proper nouns (personal names and place names)
- Fully vocalized and unvocalized Arabic script
- Accurate phonemic/phonetic transcriptions and transliteration
- All wordforms are cross-referenced to their lemma

Samples

ARAB_V	ARAB_BW	LEMMA_V	POS	GEN	NUM	NPG
بِيتْ	biyto	بِيتْ	N	M	S	000
اِلْبِيتْ	Ailobiyto	بِيتْ	N	M	S	000
بِيتِي	biytiy	بِيتْ	N	M	S	S1C
بِيتَكْ	biytako	بِيتْ	N	M	S	S2M
بِيتِكْ	biytiko	بِيتْ	N	M	S	S2F
بِيتُو	biytuw	بِيتْ	N	M	S	S3M

بِيتْهَا	biytohaA	بِيتْ	N	M	S	S3F
بِيتْنَا	biytonaA	بِيتْ	N	M	S	P1C
بِيتْكُو	biytokuw	بِيتْ	N	M	S	P2C
بِيتْهُمْ	biytohumo	بِيتْ	N	M	S	P3C

Table 2: Egyptian Arabic noun with possessive pronouns

TENSE	PRONOUN	VAR_ID	ARAB_V	ARAB_BW
Future	هُوِّ	01	رَحْ يِكْتِبْ	raHo yikotibo
Future	هُوِّ	02	رَايِحْ يِكْتِبْ	raAyiHo yikotibo
Future	هُوِّ	03	بِدُّو يِكْتِبْ	bid~uw yikotibo
Future	هُوِّ	04	رَحْ يُكْتُبْ	raHo yukotubo
Future	هُوِّ	05	رَايِحْ يُكْتُبْ	raAyiHo yukotubo
Future	هُوِّ	06	بِدُّو يُكْتُبْ	bid~uw yukotubo

Table 3: Palestinian Arabic verb variants

The CJK Dictionary Institute

The **CJK Dictionary Institute (CJKI)** was founded in 1993. Its principal activity is the compilation of large-scale dictionary databases of proper nouns and technical terms for CJK (Chinese, Japanese, Korean) and Arabic, currently with over 50 million entries. CJKI has become the world's prime source for CJK lexical resources for the IT industry and software developers, providing high-quality comprehensive dictionary data, educational tools, and consulting services.

Based in Saitama, Japan, CJKI is headed by **Jack Halpern**, editor in chief of The Kodansha Kanji Learner's Dictionary and several other dictionaries that have become standard works for learning Japanese. Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, is a lexicographer by profession, specializing in Japanese and Chinese. His work as an editor in chief of learner's dictionaries resulted in various renowned standard reference works. Born in Germany, he has lived in various countries such as France, Brazil, and the United States, and has been a resident of Japan for over 40 years. He is an avid polyglot who speaks 12 languages.