ArabLEX

برج المعاجم

Tower of Lexicons

日中韓辭典研究所
The CJK Dictionary Institute

ArabLEX
Comprehensive Arabic Full Form Lexicon
معجم اللغة العربية الكامل

by Jack Halpern

The CJK Dictionary Institute (CJKI) is pleased to announce the release of the Arabic Full Form Lexicon, or *ArabLEX.* Covering approximately **530 million entries,** this is the most comprehensive Arabic computational lexicon ever created. *Full form* means that it includes all inflected forms. It covers not only general vocabulary, but also, for the first time, fully inflected proper nouns.

ArabLEX is, quite literally, the ultimate resource for Arabic NLP and AI, ideally suited for such applications as morphological analysis, machine translation, speech technology, deep learning, and cybersecurity. No other Arabic lexicon comes close to it the scope, coverage and comprehensiveness.

ArabLEX is rich in morphological, grammatical, phonological, and orthographical attributes (currently about 30). In addition, it maps all unvocalized forms to their vocalized counterparts and to the lemma, and provides precise phonemic and phonetic transcriptions. These features can significantly contribute to the training of language models for NLP and AI applications.

The **Burj Khalifa** (*Alburj*), soaring at 832 meters above the skies of Dubai, is the tallest structure ever created. Similarly, *ArabLEX* is the largest Arabic lexicon

ever created. We thus chose Alburj as an emblem to epitomizes the essence of *ArabLEX.*

This unparalleled computational lexicon, the fruit of nearly a decade of intense development and validation, is now available to the NLP, AI, and cybersecurity communities for research and product development

1. Full Form Lexicon

1.1 What is a Full Form Lexicon?

A *full form lexicon* is a computational lexicon that contains all inflected, conjugated, declined, and cliticized forms that occur in a language (referred to as *wordforms*). Unlike ordinary dictionaries, which include only the canonical forms (base lexemes), a full form lexicon includes all wordforms. For example, the full set of wordforms for *eat* includes *eating*, *eaten* and *ate*, while for *boy* it includes *boys*, *boy's* and *boys'*. Arabic morphology is more complicated. Adding the *proclitics* وَ *wa* 'and' لِ li 'to' and the *enclitic* اتِهِمَا *tíhimạ* to the stem كَاتِبٌ kātibun 'writer' yields such a complex form as وَلِكَاتِبَاتِهِمَا *walikạtibātíhimạ* 'and to the two female writers'.

In English, the number of such forms is quite limited, but an Arabic word can have thousands of inflected and cliticized forms. For example, the verb كَتَبَ *kataba* has about 7250 forms (by comparison Japanese has about 2500), whereas the noun كَاتِب *kắtib* has about 5270 forms. As a result, the number of entries in *ArabLEX* reaches about 530 million.

1.2. Why a full form lexicon?

Traditionally, MT and other NLP applications have been (and some still are) based on rules (RBMT) or on statistical models (SMT). Recently, neural machine translation (NMT) is becoming the norm. Despite of the significant improvements that NMT has brought about, this new technology still has some

shortcomings, such as the handling of proper nouns and multiword expressions (MWE), as described in Halpern's papers on large-scale lexical resources [1] and MWEs [8]. An important issue in Arabic speech technology is the numerous complex morphological forms, like وَلِكَاتِبَاتِهِمَا *walikatibātihimā,* and the high level of orthographic ambiguity (due to the lack of vowels, as in ولكاتباتهما).

A full form lexicon can significantly contribute to the quality of Arabic MT and speech technology (both synthesis and recognition) by mapping unvocalized to vocalized forms, by providing detailed morphological information, and by providing phonemic/phonetic transcriptions for all wordforms.

2. Arabic Full Form Lexicon (*ArabLEX*)

ArabLEX is a full form Arabic lexicon that provides comprehensive coverage for inflected, conjugated and cliticized forms, and includes a rich set of attributes for natural language processing.

2.1 Distinctive features

Various features of *ArabLEX* offer special benefits to developers of Arabic NLP and AI applications, especially speech technology and machine translation.

- Created by a team of specialists in Arabic morphology and computational lexicography
- The first release in early 2021 covers about **530 million** full form entries, including proper nouns.
- Includes all inflected, conjugated, declined, and cliticized wordforms, including plurals, dual, feminine, case endings, conjugated forms, as well as proclitics, enclitics, stems and roots.
- Unvocalized and precisely fully vocalized Arabic
- Accurate phonemic transcriptions and IPA for all entries

- Millions of orthographic variants for both vocalized and unvocalized Arabic
- A large variety of grammatical codes include part-of-speech, person, gender, and case codes -- currently about 30 attributes for each entry.
- All wordforms are cross-referenced to their lemma (canonical form)
- Constantly maintained and expanded

2.2. Modules and Subsets

The full set of *ArabLEX* consists of the following major four modules and several submodules

The full version of *ArabLEX* consists of four major modules:

DAG	Database of Arabic General Vocabuary	83 million entries
DAN	Database of Arabic Names	218 million entries
DAF	Database of Arabic Foreign Names	226 million entries
DAP	Database of Arabic Place Names	6 million entries

Each module can be provided in four different subsets for specific applications. DAX represents any of the four *ArabLEX* modules. Each of these modes can be fully customized to specific requirements.

DAX-XP	Excludes proclitics, which are ten times as numerous as enclitics and may be unnecessary
DAX-PH	Includes phonemic and phonetic attributes fine tuned for speech technology, such as IPA, SAMPA, vocalization, and CARS, a specially designed phonemic transcription.
DAX-XC	Excludes all clitics, inflections and declensions (canonical forms) but can includes plural and dual forms.

DAX-WL	A bare bone wordlist, with or without inflections, and a minimum set of attributes, such POS codes.

2.3 Compilation History

For about a decade, our team of computational lexicographers and language specialists have been engaged in the development of full form lexicons for Spanish, Arabic, and Japanese. To this end, we analyzed the grammar and morphology of these languages in great depth, to a degree well beyond comprehensive descriptive grammars for these languages. Initially we focused on Spanish (bilingual SFULEX) and Japanese (monolingual JFULEX) full form lexicons, which have significantly contributed to MT technology, such as support for a Spanish-English MT system that achieved a "human quality" translation [7].

In the last couple of years we have intensified our efforts to expand, proofread and validate *ArabLEX*, with the aim of covering nearly all wordforms in Modern Standard Arabic, which contains millions of full form proper nouns in bilingual format.

3. Enhancing Speech Technology

The quality of Arabic TTS lags considerably behind that of the major languages. One reason is that the Arabic script is highly ambiguous. For example, كاتب can represent as many as *seven* pronunciations. In addition, the complexity of such cliticized forms as ولكاتباتهما *walikatibātihimā*, and the absence of vowels, makes Arabic TTS especially challenging.

The extreme orthographic ambiguity of Arabic has led to unacceptably high error rates. In a survey we discovered that sometimes over 50%, and even

80%, of the words in a sentence are mispronounced. The time has come to dramatically improve the intolerably low quality of Arabic TTS.

ArabLEX, which maps all unvocalized wordforms, including all cliticized forms, to their vocalized counterparts and provides phonemic and phonetic transcriptions that include precise word stress and even vowel neutralization. These features can help developers significantly enhance the quality of Arabic TTS by training ASR systems to achieve higher recognition rates. A module specifically designed for ASR developers provides phonemic and phonetic variants necessary for recognition but not for synthesis (such as *kaatibūn* for كاتبون in addition to the standard MSA *kaatibūna*).

To summarize, ArabLEX can bring the following benefits to speech technology:

- Hundreds of millions of full-form entries, including millions of proper nouns.
- Covers *all combinations* of proclitics and enclitics for inflected wordforms.
- Tens of millions of orthographic variants.
- Exhaustively lists alternative pronunciations for orthographical disambiguation
- Future versions will provide 'importance flags' to help indicate the most likely alternative.
- Highly accurate phonemic transcriptions for all wordforms, including precise stress and vowel neutralization.

3.1 Orthographical ambiguity

One reason that Arabic speech technology lags behind is that the Arabic script is highly ambiguous. Words are often written as a string of consonants with no indication of vowels. For example, كاتب can represent as many as *seven*

pronunciations: *kāatib*, *kātibun*, *kātibin*, *kātaba*, *kātibi*, *kātiba* and *kātibu*. Many other characteristics of the Arabic script contribute to a high level of orthographic ambiguity, as described in Halpern's paper on Arabic named entities [5].

The morphological complexity of such cliticized forms as ولكاتباتهمـا *walikạtibātíhimạ*, and the absence of vowel diacritics, makes Arabic TTS especially challenging. That is, determining the morphological composition of such forms, and the correct vowels for such consonants as ت in ولكاتباتهما often requires morphological, semantic and contextual analysis which tax the capabilities of state-of-the-art speech technology.

3.2 Improving TTS accuracy

#The extreme orthographic ambiguity of Arabic has led to unacceptably high error rates, even by the TTS systems offered by major players such as Google, Apple and Microsoft. Our institute has conducted a survey, the TTS Survey Report [9], to determine the scope of this problem. Surprisingly, we discovered that it is not unusual for over 50%, and even 80%, of the words in a sentence to be mispronounced, and that there is a trend for cliticized words to be incorrectly pronounced. For example, the cliticized word وَلِلْكَاتِبِينَ, correctly pronounced *walilkạtibīna,* is mispronounced as *walilkātibáyna.*

The tables in that report show that he error rate of Arabic TTS is unacceptably high. Such a high error rate would be unthinkable in the other major world languages. Another issue is **prosody** (stress and intonation) and **vowel neutralization** (e.g. نا *nạ* is written as a long vowel in أنا but is shortened in actual pronunciation to *na*). This is a complex issue, described in detail in Halpern's paper on Arabic stress [4].

Speech synthesis, speech recognition and prosody in current Arabic speech technology are, on the whole, inaccurate, unnatural and often unpleasant to the ear. The time has come for developers to make serious efforts to make dramatic improvements.

3.3 Improving ASR accuracy

ArabLEX includes features specifically designed to support automatic speech recognition (ASR). For speech synthesis (TTS), it is only necessary to *generate* one accurate pronunciation. For example, كاتبون 'writers' in standard Arabic is pronounced *kaṯibūna,* but for ASR it is also necessary to *recognize* the less formal variant pronunciation *kaṯibūn* Similarly, the standard pronunciation of أكتب 'I write' is *ʾáktubu,* but the final vowel is often omitted and it is pronounced *ʾáktub.*

The above alternatives are on a *phonemic* level. That is, the phoneme /na/ is being replaced by the phoneme /n/ as a result of vowel omission. There are also variations on the *phonetic* level; that is, certain phonemes have regional allophones. For example, ج in such words as جمل *jamal* is pronounced [gɛ̈mɛl] in Egypt, [ʤɛ̈mɛ̈l] in the Gulf region, and [ʒɛ̈mɛ̈l] in the Levant. It is important top note that this does not refer to the local dialects in those regions, but to a *regional varieties* of Modern Standard Arabic (MSA).

Thus *ArabLEX* not only represents ج in the standard IPA [ʤ] for TTS, it also lists the regional [ʒ] and [g] for ASR training. The goal is to enable the recognition of these allophones, but not to generate them.

3.4 Benefits to speech technology

One of the key components for training speech technology systems is the *pronunciation dictionary.* A major feature of *ArabLEX* is that it can serve as an extremely comprehensive pronunciation dictionary.

ArabLEX not only maps all unvocalized forms (including all cliticized forms) to their vocalized counterparts and to their lemmas, but also provides precise **phonemic transcriptions** (CARS system) [5] and **phonetic transcriptions** (IPA) that includes precise word stress and vowel neutralization for each entry. For example, in the IPA *wɛ̈likɛ̈ːˈtibikumɛ(ˑ*), the stressed syllable is indicated by (ˈ) (U+0C28), while (ˑ) (U+02D1) indicates that the final ɛ is neutralized vowel of optional half length. These features can help developers significantly enhance the quality of Arabic TTS, and can be used in training ASR systems to achieve higher recognition rates.

To summarize, *ArabLEX* can bring the following benefits to speech technology:

- Covers approximately **530 million entries,** including millions of proper nouns.
- Covers *all combinations* of proclitics and enclitics for inflected wordforms (mostly verbs, nouns, adjectives and proper nouns).
- Tens of millions of orthographic variants for all wordforms.
- Provides an exhaustive list of alternative pronunciations of identical unvocalized strings to enable orthographical disambiguation (e.g. six alternatives for كاتباتك).
- Future versions will provide 'importance flags' to help determine the most likely alternative.
- Highly accurate phonemic transcriptions for all wordforms, including precise stress and vowel neutralization.

- Phonetic transcriptions (IPA) indicate the correct allophonic variants in context as well as regional variants for ASR.

4. Enhancing Machine Translation

Although neural machine translation (NMT) has achieved dramatic improvements in translation quality, it does have some shortcomings, as pointed out by Philipp Koehn [3] and in Halpern's paper [1].

Some issues in Arabic MT are (1) the high orthographic ambiguity, (2) the morphological complexity (forms like ولكاتباتهما are difficult to analyze), (3) the recognition of named entities (which are often cliticized), and (4) the large number of wordforms for Arabic nouns and verbs.

ArabLEX can significantly enhance the translation accuracy of Arabic MT. Not only can it be integrated into NMT systems to provide comprehensive coverage of cliticized forms, but it can also be be used as a special kind of corpus to train the language model and enable more accurate morphological, syntactic, and semantic analysis.

When NMT first appeared, it was believed that lexicons could not be integrated into NMT systems. Later it was shown that it is technically possible to do so by regarding a lexicon as a kind of sentence-aligned, parallel corpus and assigning a higher probability to lexicon lookup results so as to override the results of the normal NMT algorithms. By using such techniques, it should therefore be possible to integrate *ArabLEX* into Arabic NMT systems [1].

5. How *ArabLEX* Works

Let us demonstrate the broad scope of the information that *ArabLEX* provides on the morphology, phonology, grammar, and orthography of Arabic words and

their thousands of inflected and cliticized forms. The stem كَاتِب *kātib* 'writer', for example, combines with the proclitics وَ *wa* and لِ *li* and the enclitic اتِهِمَا *tihimą* to yield وَلِكَاتِبَاتِهِمَا *walikątibātihimą*.

Grammatical information

Data field	Value
Full form	وَلِكَاتِبِكُمَا
Lemma	كَاتِبٌ
Stem	كَاتِب
Gender	C
Case	GEN
Number	D
Person	2
Definiteness	D
Root	ك-ت-ب

Phonological information

Data Field	Value
Unvocalized	محمد
Vocalized	مُحَمَّدٌ
Phonemic	*muhammadun*
Phonetic	mu'ħɛmmɛdun
Transliterated	muham~dN

Morphological information

Data Field	Value	Transcription
Full form	وَلِكَاتِبِكُمَا	*walikātibíkumą*
Lemma	كَاتِبٌ	*kātibun*
Stem	كَاتِب	*kātib*
Proclitic	وَلِ	*wali*
Enclitic	كُمَا	*(i)kúmą*
Root	ك-ت-ب	*k-t-b*

Orthographical information

Data Field	Value	Transcription
Variant 1	ألكسندرة	*ʾaliksándara*
Variant 2	الكسندرة	*ʾaliksándara*
Variant 3	ألكسندره	*ʾaliksándara*
Variant 4	الكسندره	*ʾaliksándara*

5.1 Grammatical information

This includes gender codes, number codes, case ending codes, persons codes, the stem, the state (definiteness), and the lemma. For وَلِكَاتِبَاتِهِمَا, *ArabLEX* provides as the following grammatical information.

Table 1: Grammatical information

Data field	Value	Description
Full form	وَلِكَاتِبِكُمَا	*walikātibíkumạ*
Lemma	كَاتِبٌ	*kātibun*
Stem	كَاتِب	*kātib*
Gender	C	common gender (masculine & feminine)
Case	GEN	genitive case
Number	D	dual
Person	2	second person
State	D	definite, indefinite or construct state
Root	ك-ت-ب	the triliteral root

Abundant grammatical information is useful for morphological analysis, orthographic disambiguation, semantic analysis, and pedagogical applications.

5.2 Phonological information

This includes full vocalization and phonemic transcription for all unvocalized Arabic, including the full form headword, the lemma, the stem, proclitics, and enclitics. The headword is also given in IPA, a precise phonetic transcription that includes word stress. For example, for each full form Arabic headword the following fields are given:

Table 2: Phonological information

Data field	Value	Description
Unvocalized	محمد	Unvocalized Arabic as it actually occurs
Vocalized	مُحَمَّدٌ	precise and full vocalization
Phonemic	*muhammadun*	accurate phonemic transcription in CARS system with vowel neutralization
Phonetic	muˈħɛmmɛdun	phonetic transcription in IPA (or optionally SAMPA) with word stress
Transliterated	muham~dN	orthographic transliteration in the Buckwalter system

The phonemic and phonetic transcriptions provide precise information that is useful for training speech technology systems, both TTS and ASR.

5.3 Morphological information

This includes the lemma, stem, proclitics (prefixes) and enclitics (suffixes) for each full form headword, as shown below.

Table 3: Morphological information

Data field	Value	Transcription
Full form	وَلِكَاتِبِكُمَا	*walikātibíkumạ*

Lemma	كَاتِبٌ	*kātibun*
Stem	كَاتِب	*kātib*
Proclitic	وَلِ	*wali*
Enclitic	كُمَا	*(i)kúmạ*
Root	ك-ت-ب	k-t-b

The morphological information is useful for morphological analysis, semantic analysis, lemmatization, decliticization, verb conjugation, and dictionary lookup.

5.4 Orthographical information

This includes the vocalized and unvocalized headwords as well as their orthographic variants, as shown below.

Table 4: Orthographical information

Data field	Value	Transcription
Vocalized headword	أَلْكَاتِبٌ	*ʾalkātibun*
Unvocalized headword	آلكاتب	*ʾalkātibun*
Unvocalized variant	أالكاتب	*ʾalkātibun*

The orthographical information is useful for word/entity recognition, word/entity extraction, normalization, and dictionary lookup.

5.5 Orthographical disambiguation

A central issue in Arabic NLP applications, especially in speech technology, is identifying which of the possible wordforms an Arabic string like كاتباتـك represents. This can represent six wordforms shown in the table below, each

with a different meaning and different morphological or syntactic function. We will refer to this process as *orthographical disambiguation.*

Table 5: Orthographical disambiguation

POS	ARAB_V	ARAB_U	ARAB_T	GEN	NUM	CASE	PER	DEF
N	كَاتِبَاتُكَ	كاتباتك	kạtibātuka	F	P	NOM	2SM	d
N	كَاتِبَاتُكِ	كاتباتك	kạtibātuki	F	P	NOM	2SF	d
N	كَاتِبَاتِكَ	كاتباتك	kạtibātika	F	P	GEN	2SM	d
N	كَاتِبَاتِكِ	كاتباتك	kạtibātiki	F	P	GEN	2SF	d
N	كَاتِبَاتِكَ	كاتباتك	kạtibātika	F	P	ACU	2SM	d
N	كَاتِبَاتِكِ	كاتباتك	kạtibātiki	F	P	ACU	2SF	d

The rich set of grammatical attributes shown above (and the morphological attributes not shown) can help train the language model to correctly disambiguate such ambiguous forms. That is, they provide the grammatical and morphological context in which كاتباتك can occur, helping determine the specific correct wordform for that context, and thus the correct pronunciation. For example, the attributes for كَاتِبَاتِكِ show that it refers to plural female writers in the definite state, who belong to second person singular feminine in the genitive case, which helps to determine the correct pronunciation of *kạtibātiki* in training speech technology models.

6. Conclusions

ArabLEX provides a rich set of morphological and phonological features. It brings the following benefits to speech technology and MT:

- Enhance the quality of MT, NLP and AI applications.
- Full support for accurate morphological analysis, including stemming, lemmatization, segmentation and tokenization.

- Supplements corpora for training speech technology models.
- Improved accuracy of word and entity recognition and extraction.
- Support for query processing in information retrieval applications.
- Support for automatic conjugators for pedagogical and NLP applications.
- Part-of-speech analysis and POS tagging.
- Accurate determination of the root for each wordform.

In summary, *ArabLEX* aims to serve as the ultimate resource for Arabic natural language processing.

References

[1] Halpern, Jack. Very Large-Scale Lexical Resources to Enhance Chinese and Japanese Machine Translation

[2] Habash, Nizar Y. Introduction to Arabic Natural Language Processing

[3] Koehn, Philipp. The State of Neural Machine Translation (NMT)

[4] Halpern, Jack. Word Stress and Vowel Neutralization in Modern Standard Arabic

[5] Halpern, Jack. CJKI Arabic Romanization System

[6] Halpern, Jack. Lexicon-Driven Approach to the Recognition of Arabic Named Entities

[7] Carbonell, Jaime; Klein, Steve; Miller, David; Steinbaum, Michael. Context-Based Machine Translation

[8] Wajdan Algihab, Imam Muhammad bin Saud Islamic University. Arabic Speech Recognition with Deep Learning: A Review

[9] Halpern, Jack. TTS Survey Report

APPENDIX 1: DATA FIELDS

Table 6: Basic data fields

No	Field	Value	Field Description
1	IDENTIFIER	05373	unique ID for Arabic headword (full form) or variant thereof
2	SUBID	00	identifies variants of headword
3	POS	N	part of speech code
4	ARAB_V	وَلِكَاتِبِكُمَا	vocalized Arabic headword (full form)
5	ARAB_U	ولكاتبكما	unvocalized Arabic headword (full form)
6	ARAB_T	walikātibíkumạ	Arabic headword in phonemic transcription (full form) in CARS system (including vowel neutralization)
7	LEMMA_V	كَاتِبٌ	lemma in vocalized Arabic
8	LEMMA_U	كاتب	lemma in unvocalized Arabic
9	LEMMA_T	kắtibun	lemma in phonemic transcription in CARS
10	GEN	M	gender code for stem:masculine

11	NUM	S	number code for stem: singular
12	CASE	GEN	case ending code for nouns and adjectives: genitive
13	PER	2DC	person code for full form (ARABIC_V): second person dual common gender
14	DEF	D	Definite, indefinite or construct state
15	TYPE	N/A	code for verb conjugation (for verbs only)
16	TENSE	N/A	code for verb tense (for verbs only)

Table 7: Advanced data fields

No	Field	Value	Field Description
17	PROC_V	وَلِ	prefix or proclitic in vocalized Arabic
18	PROC_U	ول	prefix or proclitic in unvocalized Arabic
19	PROC_T	wáli	prefix or proclitic in phonemic transcription
20	STEM_V	كَاتِب	stem in vocalized Arabic
21	STEM_U	كاتب	stem in unvocalized Arabic
22	STEM_T	kā́tib	stem in phonemic transcription in CARS

23	ENC_V	كُمَا	suffix or enclitic in vocalized Arabic
24	ENC_U	كما	suffix or enclitic in unvocalized Arabic
25	ENC_T	ikúmą	suffix or enclitic in phonemic transcription in CARS
26	ROOT	ك-ت-ب	the root of each headword
27	TRANSLIT	walikaAtibikumaA	graphemic transliteration in the Buckwalter system
28	IPA	wɛ̈likɛ̈:'tibikumɛ(˙)	phonetic transcription of full-form headword in IPA, including word stress and vowel neutralization

APPENDIX 2: DATA SAMPLE

Below is a small subset of *ArabLEX* for the noun كَاتِبٌ *kātibun,* which contains a total of about 5000 wordforms. The full sample can be found at *ArabLEX* _sample.xls.

Table 8: Sample for nouns (fields 3 to 9)

POS	ARAB_V	ARAB_U	ARAB_T	LEMMA_V	LEMMA_U	LEMMA_T
N	فَلِكَاتِبِكُمَا	فلكاتبكما	falikātibíkumạ	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمْ	فلكاتبهم	falikātíbihim	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِنَّ	فلكاتبهن	falikātibihínna	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمَا	فلكاتبهما	falikātibíhimạ	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمَا	فلكاتبهما	falikātibíhimạ	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِي	فلكاتبي	falikātibi	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبَيْنِ	فلكاتبين	falikātibáyni	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبَيَّ	فلكاتبي	falikātibáyya	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبَيْكَ	فلكاتبيك	falikātibáyka	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبَيْكِ	فلكاتبيك	falikātibáyki	كَاتِبٌ	كاتب	kā́tibun

Table 9: Sample for nouns (fields 10-16)

ARAB_V	GEN	NUM	CASE	PER	DEF	TYPE	TENSE
فَلِكَاتِبِكُمَا	M	S	GEN	2DC	d	-	-
فَلِكَاتِبِهِمْ	M	S	GEN	3PM	d	-	-
فَلِكَاتِبِهِنَّ	M	S	GEN	3PF	d	-	-
فَلِكَاتِبِهِمَا	M	S	GEN	3DM	d	-	-
فَلِكَاتِبِهِمَا	M	S	GEN	3DF	d	-	-
فَلِكَاتِبِي	M	S	ACU	1SC	d	-	-
فَلِكَاتِبَيْنِ	M	D	GEN	0	D	-	-
فَلِكَاتِبَيَّ	M	D	GEN	1SC	d	-	-
فَلِكَاتِبَيْكَ	M	D	GEN	2SM	d	-	-
فَلِكَاتِبَيْكِ	M	D	GEN	2SF	d	-	-

Table 10: Sample for nouns (fields 17-22)

ARAB_V	PROC_V	PROC_U	PROC_T	STEM_V	STEM_U	STEM_T
فَلِكَاتِبِكُمَا	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبِهِمْ	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبِهِنَّ	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib

فَلِكَاتِبِهِمَا	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبِهِمَا	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبِي	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبَيْنِ	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبَيَّ	فَلِ	فل	fáli	كَاتِب	كاتب	kắtib
فَلِكَاتِبَيْكَ	فَلِ	فل	fáli	كَاتِب	كاتب	kātib
فَلِكَاتِبَيْكِ	فَلِ	فل	fáli	كَاتِب	كاتب	kātib

Table 11: Sample for nouns (fields 23-28)

ARAB_V	ENC_V	ENC_U	ENC_T	ROOT	BW	IPA
فَلِكَاتِبِكُمَا	كُمَا	كما	ikúmą	ك-ت-ب	falikaAtibikumaA	
فَلِكَاتِبِهِمْ	هِمْ	هم	ihim	ك-ت-ب	falikaAtibihimo	
فَلِكَاتِبِهِنَّ	هِنَّ	هن	íhínna	ك-ت-ب	falikaAtibihin~a	
فَلِكَاتِبِهِمَا	هِمَا	هما	íhímą	ك-ت-ب	falikaAtibihimaA	
فَلِكَاتِبِهِمَا	هِمَا	هما	íhímą	ك-ت-ب	falikaAtibihimaA	
فَلِكَاتِبِي	ي	ي	i	ك-ت-ب	falikaAtibiy	
فَلِكَاتِبَيْنِ	يْنَ	ين	áyni	ك-ت-ب	falikaAtibayoni	

فَلِكَاتِبَيَّ	يَّ	ي	áyya	ك-ت-ب	falikaAtibay~a	
فَلِكَاتِبَيْكَ	يْكَ	يك	áyka	ك-ت-ب	falikaAtibayoka	
فَلِكَاتِبَيْكِ	يْكِ	يك	áyki	ك-ت-ب	falikaAtibayoki	