AFULEX

Comprehensive Arabic Full Form Lexicon

The CJK Dictionary Institute
by Jack Halpern

Abstract

The CJK Dictionary Institute (CJKI) is pleased to announce the release of the **Arabic Full Form Lexicon** (AFULEX), covering approximately **130 million** (eventually 250 million) inflected, conjugated and cliticized wordforms. AFULEX is not only comprehensive in coverage, it is also rich in morphological, grammatical, phonological, and orthographical attributes. 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 speech technology (both synthesis and recognition) and machine translation.

This unparalleled computational lexicon, the fruit of nearly a decade of intense development and validation, is now available to the NLP community for research and product development. It aims to serve as the ultimate resource for Arabic natural language processing.

1. What is a Full Form Lexicon?

1.1 What is a full form lexicon?

A *full form lexicon* is a computational lexicon that contains all inflected, conjugated, and cliticized forms that occur in a language (referred to as *wordforms)*. Unlike ordinary dictionaries, which include only the *canonical* forms or *lemmas* (base lexemes), a full form lexicon includes all wordforms. For example, the full set of wordforms for the verb *eat* includes *eating*, *eaten* and *ate,* while for the noun *boy* it includes *boys, boy's* and *boys'.* Arabic nouns, on the other hand, as pointed out by the renowned linguist Nizar Habash, are "far more complex and idiosyncratic" [2]. Attaching the *proclitics* وَ *wa* 'and' لِ *li* 'to' and the *enclitic* اتِهِمَا *tíhima̱* to the stem كَاتِب *kātibun* 'writer' yields such a complex form as وَلِكَاتِبَاتِهِمَا *walika̱tibātíhima̱* '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 *kātib* has about 5270 forms. As a result, the number of entries in the Basic Edition of AFULEX reaches 130 million, and the Expanded Edition is expected to reach about 250 million entries.

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 *walika̱tibātíhima̱,* 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.

2.1 Arabic Full Form Lexicon (AFULEX)

AFULEX 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 AFULEX offer special benefits to developers of Arabic NLP applications, especially speech technology and machine translation.

- Created by a team of specialists in Arabic morphology and computational lexicography
- The **Basic Edition** contains some **130 million entries,** to be expanded to about **250 million entries** in the Expanded Edition
- Includes all inflected, conjugated, and cliticized wordforms
- Wordforms include plurals, dual, feminine, case endings, conjugated forms, and all proclitics and enclitics
- Unvocalized mapped to precisely fully vocalized Arabic
- Highly accurate phonemic transcription for all entries
- Millions of orthographic variants for both vocalized and unvocalized Arabic
- Various grammatical codes include part-of-speech, person, gender, and case codes
- The **Deluxe Edition** includesproclitics, enclitics, stems, accurate IPA and word stress
- All wordforms are cross-referenced to the lemma (canonical form)
- Provides allophones of regional varieties of MSA to support ASR
- Constantly maintained and expanded

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 the Basic Edition of AFULEX, with the aim of covering nearly all wordforms in Modern Standard Arabic. The Expanded Edition will contain millions of full form proper nouns in bilingual format.

3. Enhancing Speech Technology

The recent advances in deep learning and neural network technology have dramatically improved speech technology [8]. Although the quality of Arabic speech technology has been steadily improving, it still lags significantly behind the other major languages, such as Chinese and Japanese.

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 ولكاتباتهمـــا *walika̱tibātíhima̱*, 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 to determine the scope of this problem, some of the results of which are reported in Appendix 3. 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 وَلِلْكَـــاتِبِينَ, correctly pronounced *walilka̱tibīna,* is mispronounced as *walilkātibáyna.*

Appendix 3 shows that the 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. *na̱* is a long vowel نا shortened in actual pronunciation)**.** 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

The Expanded Version of AFULEX will include a module 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 *ka̱tibūna,* but for ASR it is also necessary to *recognize* the less formal variant pronunciation *ka̱tibūn* 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, [d̠ʒɛ̈mɛ̈l] 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 AFULEX not only represents ج in the standard IPA [d̠ʒ] 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 AFULEX is that it can serve as an extremely comprehensive pronunciation dictionary.

AFULEX 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, AFULEX can bring the following benefits to speech technology:

- The Basic Edition covers approximately 130 million entries, while the Expanded Edition will probably cover about 250 millions 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, even in this era of neural-based algorithms (NMT), 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. These issues are described in more detail in Halpern's paper on Arabic named entities [5].

AFULEX can significantly enhance the translation accuracy of Arabic MT. Not only can it be directly integrated into NMT systems to provide comprehensive coverage of cliticized forms, but it can also 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 AFULEX into Arabic NMT systems [1].

5. How AFULEX works

Let us demonstrate the broad scope of the information that AFULEX 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 *tíhima̱* to yield وَلِكَاتِبَاتِهِمَا *walika̱tibātíhima̱*. The enclitic اتِهِمَا indicates the third person dual feminine in the genitive case, the proclitic وَ means 'and' and the proclitic لِ means 'for, to; in order to', so that the full form وَلِكَاتِبَاتِهِمَا means something like 'and to the two female writers'. Below is a description of how such information is presented.

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, AFULEX provides as the following grammatical information.

Table 1: Grammatical information

Data field	Value	Description
Full form	وَلِكَاتِبِكُمَا	*walikātibíkuma̱*
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	مُحَمّدٌ	precise and full vocalization
Phonemic	*muhammadun*	accurate phonemic transcription in CARS system with vowel neutralization
Phonetic	muˈħɛ̈mmɛ̈dun	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	*walikātibíkuma̱*
Lemma	كَاتِبٌ	*kātibun*
Stem	كَاتِب	*kātib*
Proclitic	وَلِ	*wali*
Enclitic	كُمَاِ	*(i)kúma̱*
Root	ك-ت-ب	k-t-b

The morphological information is useful for morphological analysis, semantic analysis, lemmatization, decliticization, verb conjugation, and dictionary lookup. Thus AFULEX provides full support for morphological analysis, including such operations as

Operation	Vocalized	Unvocalized
Full form	وَلْيَكْتُبُوكُمَا	وليكتبوكما
Lemmatization	كَتَبَ	كتب
Segmentation	وَ + لْ + يَكْتُبُو + كُمَا	و + ل + يكتبو + كما
Tokenization	وَ + لْ + يَكْتُبُوا + كُمَا	و + ل + يكتبوا + كما
Stemming	كَتَب, يَكْتُبُوا	يكتبوا, كتب
Root extraction	-	ك-ت-ب

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	*ʾa̲lkātibun*
Unvocalized headword	آلكاتب	*ʾa̲lkātibun*
Unvocalized variant	ألكاتب	*ʾa̲lkātibun*

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 كاتباتك 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	ka̲tibātuka	F	P	NOM	2SM	d
N	كَاتِبَاتُكِ	كاتباتك	ka̲tibātuki	F	P	NOM	2SF	d
N	كَاتِبَاتِكَ	كاتباتك	ka̲tibātika	F	P	GEN	2SM	d
N	كَاتِبَاتِكِ	كاتباتك	ka̲tibātiki	F	P	GEN	2SF	d
N	كَاتِبَاتِكَ	كاتباتك	ka̲tibātika	F	P	ACU	2SM	d
N	كَاتِبَاتِكِ	كاتباتك	ka̲tibātiki	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 *ka̱tibātiki* in training speech technology models.

6. Conclusions

AFULEX is a comprehensive Arabic lexical database that provides a rich set of grammatical, morphological and phonological features. It brings the following benefits to NLP, especially speech technology and machine translation.

- Enhance the quality NLP applications, especially MT, speech technology and morphological analysis.
- Full support for accurate morphological analysis, including stemming, lemmatization, segmentation and tokenization.
- Supplements corpora in training speech TTS and ARS systems/
- 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 of each wordform.

In summary, AFULEX 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

APPENDIX 1: DATA FIELDS

Table 6: Data fields in Basic Edition

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**	walikātibíkuma̱	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**	kā́tibun	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: Data fields in Deluxe Edition

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	ikúma̱	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	wɛ̈likɛ̈ː'tibikumɛ(ˑ)	phonetic transcription of full-form headword in IPA, including word stress and vowel neutralization

APPENDIX 2: DATA SAMPLE

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

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	falikātibíkuma̱	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمْ	فلكاتبهم	falikātíbihim	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِنَّ	فلكاتبهن	falikātibihínna	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمَا	فلكاتبهما	falikātibíhima̱	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِهِمَا	فلكاتبهما	falikātibíhima̱	كَاتِبٌ	كاتب	kā́tibun
N	فَلِكَاتِبِي	فلكاتبي	falikā́tibi̱	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	kā□tib
فَلِكَاتِبِهِمْ	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبِهِنَّ	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبِهِمَا	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبِهِمَا	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبِي	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبَيْنِ	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
فَلِكَاتِبَيَّ	فَلِ	فل	fáli	كَاتِب	كاتب	kā□tib
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	ikúma□	ك-ت-ب	falikaAtibikumaA	
فَلِكَاتِبِهِمْ	هِمْ	هم	ihim	ك-ت-ب	falikaAtibihimo	
فَلِكَاتِبِهِنَّ	هِنَّ	هن	íhínna	ك-ت-ب	falikaAtibihin~a	

فَلِكَاتِبِهِمَا	هِمَا	هما	íhíma□	ك-ت-ب	falikaAtibihimaA	
فَلِكَاتِبِهِمَا	هِمَا	هما	íhíma□	ك-ت-ب	falikaAtibihimaA	
فَلِكَاتِبِي	ي	ي	i□	ك-ت-ب	falikaAtibiy	
فَلِكَاتِبَيْنِ	يْنِ	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	

APPENDIX 3: TTS SURVEY RESULTS

Below are the results of a survey conducted by CJKI (our institute) to compare the TTS systems of Google, Apple (iPhone) and Microsoft (Bing), showing high error rates for all three. The **Unvocalized** field is the original Arabic text, the **Vocalized** field indicates the correct pronunciation, and the **CJKI** field shows the correct pronunciation in CARS phonemic transcription [5]. The CARS transcriptions in these **Google, iOS** and **Bing** columns indicate the actual pronunciation by the three TTS engines. Mispronunciations are indicated in red, and the error rate is given in the column headers.

Table 12 and 14 are based on text c**omposed** for this survey, while tables 13 and 15 use a sentence **extracted** from the web. (It is noteworthy that the error rate for the composed text is actually much lower than for the extracted text.) Tables 12 and 13 compare the results on a word-by-word basis, whereas tables 14 and 15 compare them on a sentence-by-sentence basis, showing the context. The fact that the error rate is sometimes over 80% is surprising and unacceptable to users.

Table 12: Mispronounced Words in Composed Text

Unvocalized	Vocalized	Google (13%)	iOS (31%)	Bing (25%)	CJKI (0%)
عدد	عَدَّدَ	ɛádadu	ɛádada	ɛádada	ɛáddada
الكاتب	ٱلْكَاتِبُ	lkātibu	lkātibi	lkātibu	lkātibu
ما	مَا	ma̱	ma̱	ma̱	ma̱
قال	قَالَ	qāla	qāla	qāla	qāla
إن	إِنَّ	ʾínna	ʾínna	ʾínna	ʾínna

هؤلاء	هٰؤُلَاءِ	ha̱ʾulāʾi	ha̱ʾulāʾi	ha̱ʾulāʾi	ha̱ʾulāʾi
الحكام	ٱلْحُكَّامَ	lḥukkāmi	lḥukkāmi	lḥukkāmi	lḥukkāma
يفعلونه	يَفْعَلُونَهُ	yafεalūnahu	yafεalūnahu	yafεalūnahu	yafεalūnahu
في	فِي	fi̱	fi̱	fi̱	fi̱
الخارج	ٱلْخَارِجِ	lkhāriji	lkhārija	lkhāriji	lkhāriji
مثل	مِثْلَ	míthli	míthli	míthli	míthla
الهجمات	ٱلْهَجَمَاتِ	lhajamāti	lhajamāti	lhajamāti	lhajamāti
الإلكترونية	ٱلْإِلِكْتُرُونِيَّةِ	lʾilikturu̱níyyati	lʾilikturu̱níyyati	lʾilikturu̱níyyati	lʾilikturu̱níyyati
ومطاردة	وَمُطَارَدَةِ	wamuṭārádati	wamuṭārídati	wamuṭārídati	wamuṭārádati
المعارضين	ٱلْمُعَارِضِينَ	lmuεa̱riḍīna	lmuεa̱riḍīna	lmuεa̱riḍīna	lmuεa̱riḍīna
اللاجئين	ٱللَّاجِئِينَ	lla̱jiʾīna	lla̱jiʾīna	lla̱jiʾīna	lla̱jiʾīna
في	فِي	fi̱	fi̱	fi̱	fi̱
العواصم	ٱلْعَوَاصِمِ	lεawāṣimi	lεawāṣimi	lεawāṣimi	lεawāṣimi
الغربية	ٱلْغَرْبِيَّةِ	lgharbíyyati	lgharbíyyati	lgharbíyyati	lgharbíyyati
وللكاتبين	وَلِلْكَاتِبِينَ	walilka̱tibīna	walilkātibáyna	walilkātibáyna	walilka̱tibīna
من	مِنَ	mína	mína	mína	mína
الصحفيين	ٱلصَّحَفِيِّينَ	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna	ṣṣaḥafi̱yīna
العرب	ٱلْعَرَبِ	lεárabi	lεárabi	lεárabi	lεárabi
صرح	صَرَّحَ	ṣárraḥa	ṣáraḥa	ṣáraḥa	ṣárraḥa
بأن	بِأَنَّ	biʾánna	biʾánna	biʾánna	biʾánna

عليهم	عَلَيْهِمْ	ɛaláyhim	ɛaláyhim	ɛaláyhim	ɛaláyhim
أن	أَنْ	ʾan	ʾan	ʾan	ʾan
يكتبوا	يَكْتُبُوا	yaktúbuwu	yaktúbuwu	yaktúbuwu	yaktúbuwu
ما	مَا	ma̱	ma̱	ma̱	ma̱
تمليه	تُمْلِيهِ	tumallī̆hi	tamlī̆hi	tamlī̆hi	tumlī̆hi
عليهم	عَلَيْهِمْ	ɛalayhim	ɛalayhim	ɛalayhim	ɛalayhim
ضمائرهم	ضَمَائِرُهُمْ	ḍamāʾíruhum	ḍamāʾírihim	ḍamāʾírihim	ḍamāʾíruhum

Table 13: Mispronounced Words in Extracted Text

Unvocalized	Vocalized	Google (80%)	iOS (90%)	Bing (70%)	CJKI (0%)
الاخوات	ٱلْأَخَوَاتُ	ʾalikhwā̆tu	ʾalʾakhawā̆ti	ʾalʾakhawā̆tu	ʾalʾakhawā̆tu
المتزوجات	ٱلْمُتَزَوِّجَاتُ	lmutazawwijā̆tu	lmutazawwijā̆ti	lmutazawwijā̆ti	lmutazawwijā̆tu
اللاتي	ٱللَّاتِي	lṯi	llaṯi	llā̆ṯi	llā̆ṯi
رزقن	رُزِقْنَ	rízqin	rúzqin	rúzqin	ruzíqna
بابناء	بِأَبْنَاءَ	ba̱binā̆ʾun	bibnā̆ʾi	bibnā̆ʾi	biʾabnā̆ʾa
فليكتبن	فَلْيَكْتُبْنَ	falayiktíbna	falktíbna	falktíbna	falyaktúbna
اسمائهم	أَسْمَائَهُمْ	ʾismā̆ʾahum	smā̆ʾihim	smā̆ʾihim	ʾasmā̆ʾahum
وسبب	وَسَبَبَ	wasábaba	wasábaba	wasábaba	wasábaba
التسميه	ٱلتَّسْمِيَةِ	lttasammī̆hu	lttasammī̆hi	lttasammī̆hi	ttasmíyati
رجاء	رَجَاءًا	rajjā̆ʾan	rajā̆ʾ	rajā̆ʾ	rajā̆ʾan

Table 14: Mispronounced Sentences in Composed Text

TTS	Sentence	Error %
Unvocalized	عـدد الكـاتب مـا قـال إن هـؤلاء الحكـام يفعلـونه فـي الخـارج مثـل الهجمـات الإلكترونيـة ومطـاردة المعارضـين اللاجئيـن فـي العواصـم الغربية. وللكاتبين من الصحفيين العرب صرح بأن عليهم أن يكتبوا ما تمليه عليهم ضمائرهم	-
Vocalized	عَدَّدَ ٱلْكَاتِبُ مَـا قَـالَ إِنَّ هٰـؤُلَاءِ ٱلْحُكَّـامَ يَفْعَلُـونَهُ فِـي ٱلْخَـارِجِ مِثْـلَ ٱلْهَجَمَات ٱلْإِلِكْتُرُونِيَّـةِ وَمُطَـارَدَةَ ٱلْمُعَارِضِـينَ ٱللَّاجِئِيـنَ فِـي ٱلْعَوَاصِـمِ ٱلْغَرْبِيَّةَ. وَلِلْكَاتِبِينَ مِنَ ٱلصَّحَفِيِّينَ ٱلْعَرَبِ صَرَّحَ بِـأَنَّ عَلَيْهِـمْ أَنْ يَكْتُبُـوُا مَا تُمْلِيهِ عَلَيْهِمْ ضَمَائِرُهُمْ	0%
CJKI	εáddada lkā́tibu ma̲ qā́la ʾínna ha̲ʾulā́ʾi lḥukkā́ma yafεalū́nahu fi̲ lkhā́riji míthla lhajamā́ti lʾilikturu̲níyyati wamuṭārádati lmuεa̲riḍīna lla̲jiʾīna fi̲ lεawā́ṣimi lgharbíyyati. walilka̲tibīna mína ṣṣaḥafi̲yīna lεárabi ṣárraḥa biʾánna εaláyhim ʾan yaktúbuwu ma̲ tumlī́hi εalayhim ḍamāʾíruhum	0%
Google	εádadu lkā́tibu ma̲ qā́la ʾínna ha̲ʾulā́ʾi lḥukkā́mi yafεalū́nahu fi̲ lkhā́riji míthli lhajamā́ti lʾilikturu̲níyyati wamuṭārádati lmuεa̲riḍīna lla̲jiʾīna fi̲ lεawā́ṣimi lgharbíyyati. walilka̲tibīna mína ṣṣaḥafi̲yīna lεárabi ṣárraḥa biʾánna εaláyhim ʾan yaktúbuwu ma̲ tumallī́hi εalayhim ḍamāʾíruhum	13%
iOS	εádada lkā́tibi ma̲ qā́la ʾínna ha̲ʾulā́ʾi lḥukkā́mi yafεalū́nahu fi̲ lkhā́rija míthli lhajamā́ti lʾilikturu̲níyyati wamuṭārídati lmuεa̲riḍīna lla̲jiʾīna fi̲ lεawā́ṣimi lgharbíyyati. walilkātibáyna mína ṣṣaḥafi̲yīna lεárabi ṣáraḥa biʾánna εaláyhim ʾan yaktúbuwu ma̲ tamlī́hi εalayhim ḍamāʾírihim	31%
Bing	εádada lkā́tibu ma̲ qā́la ʾínna ha̲ʾulā́ʾi lḥukkā́mi afεalū́nahu fi̲ lkhā́riji míthli lhajamā́ti lʾilikturu̲níyyati wamuṭārídati lmuεa̲riḍīna lla̲jiʾīna fi̲ lεawā́ṣimi lgharbíyyati. walilkātibáyna mína ṣṣaḥafi̲yīna lεárabi ṣáraḥa biʾánna εaláyhim ʾan yaktúbuwu ma̲ tamlī́hi εalayhim ḍamāʾírihim	25%

Table 15: Mispronounced Sentences in Extracted Text

TTS	Sentence	Error %
Unvocalized	الاخوات المتزوجـات اللاتـي رزقـن بابنـاء فليكتبـن اسـمائهم وسـبب التسميه رجاءا	-
Vocalized	اَلْأَخَوَاتُ ٱلْمُتَزَوِّجَـاتُ ٱللَّاتِـي رُزِقْـنَ بِأَبْنَـاءَ فَلْيَكْتُبْـنَ أَسْـمَائَهُمْ وَسَـبَبَ	0%

	ٱلتَّسْمِيَةِ رَجَاءً	
CJKI	ʾalʾakhawātu lmutazawwijātu llāṯi ruzíqna biʾabnāʾa falyaktúbna ʾasmāʾahum wasábaba ttasmíyati rajāʾan	0%
Google	ʾalikhwātu lmutazawwijātu lṯi rízqin ba̱bināʾun falayiktíbna ʾismāʾahum wasábaba lttasammīhu rajjāʾan	80%
iOS	ʾalʾakhawāti lmutazawwijāti llaṯi rúzqin bibnāʾi falktíbna smāʾihim wasábaba lttasammīhi rajāʾ	90%
Bing	ʾalʾakhawātu lmutazawwijāti llāṯi rúzqin bibnāʾi falktíbna smāʾihim wasábaba lttasammīhi rajāʾ	70%