CJKI Arabic Romanization System
CARS

Jack Halpern
The CJK Dictionary Institute, Inc.
jack@cjki.org

Abstract

The **CJKI Arabic Romanization System,** or ***CARS***, is an innovative phonemic transcription system developed mainly for ease of use by learners and as an aid to linguists in analyzing the phonological structure of Modern Standard Arabic (MSA). The most important feature of CARS is its readability: its unambiguous encoding of phonological and prosodic information in an intuitive, easy to read set of symbols. The new system, which is making its debut with the appearance of *The CJKI Arabic Learner's Dictionary,* has several unique features not found elsewhere, including a symbol set that represents all Arabic phonemes unambiguously, and, for the first time, symbols indicating word stress and vowel neutralization. This paper describes how CARS is used to represent phonemic, prosodic, syllabic, grammatical and some phonetic information in a user-friendly manner.

1. Introduction

1.1 What is romanization

Romanization is using the letters of the Latin alphabet to represent a language written in a non-Roman script, such as Japanese, Chinese or Arabic. This includes both transliteration and various kinds of transcription. Much confusion surrounds these terms, with the former often being misleadingly used in the sense of the latter even in academic papers (Jaleel and Larkey, 2003). Briefly, **transliteration** refers to representing the source script (graphemes, not phonemes) with the characters of another script, as in محمد \mHmd\. **Transcription** is representing the source script in the target script in a manner that reflects pronunciation (Halpern, 2009a). This includes *phonetic transcription,* which transcribes actual speech sounds, as in [muħɛ̈mmɛ̈d] (IPA), *phonemic transcription,* which represents the phonemes of the source language, as in /muHammad/, and *popular transcription,* a conventionalized orthography that roughly represents pronunciation, as in *Mohammed, Muhammad, Moohammad, Moohamad…*and almost 200 others.

1.2 Orthographic Ambiguity

As is well known, unvocalized Arabic is highly ambiguous (Halpern, 2007). A string like كاتب can represent the following seven wordforms: كَاتِب ***kātib (noun),*** كَاتَبَ ***kātaba (noun acc. def.),*** كَاتِبٍ ***kātibin (noun acc. indef.)***, كَاتِبٌ ***kātibun (noun nom.)***, كَاتِبَ ***kātiba (verb)***, كَاتِبِ ***kātibi (noun construct gen.)*** and كَاتِبُ ***kātibu (noun construct nom.)***. But even fully vocalized Arabic, which includes such rare diacritics as *dagger ʾalif* (هٰذَا ***hā́dha̱***) and ***waṣla*** (ٱ), is not a truly phonemic script since it does not represent all the phonemes unambiguously without resort to rules. Even if all the diacritics are carefully added (which is rarely the case), they still do not convey enough information to enable the reader to pronounce correctly without a knowledge of orthographic rules.

There are various reasons for the ambiguity of vocalized Arabic. On the one hand, some letters are not pronounced at all. For example:

1. ٱ (*ʾalif waṣla*) in مِنَ ٱلْمُمْكِن *mina‿lmúmkini* ‘possible’ (normally omitted even in vocalized Arabic)
2. ل (*lām*) in assimilated articles as in اَلدَّخْلُ *addákhlu* 'income'
3. final long vowels, as in أَنَا *ʾána̱* ‘I’, هٰذَا *hā́dha̱* 'this' (final *ā* is shortened)
4. final double consonants, as in حُبّ *ḥuḇ* ‘love’ (final double /b/ becomes single)
5. *ʾalif* in certain verb forms, as in كَتَبُوا *kátabu* ‘they wrote’

On the other hand, such features as word stress and vowel neutralization, both of great importance to correct pronunciation, are not indicated in the orthography at all. Looking at a word like كَتَبَ *kataba* ‘he wrote’, there is no way to determine which syllable is stressed -- *kátaba, katába* or *katabá* -- unless one knows a set of fairly complex rules and exceptions. Naturally, there is also no way to know from a fully vocalized Arabic word, such as هٰذَا *hā́dha̱* 'this' and عَلَى *εála̱* 'on', that the final long vowel is neutralized (shortened).

1.3 Why romanization

Since one of the primary goals of CARS is to aid learners, let us pose a fundamental question: do learners of Arabic even need a romanization system? The answer is a definitive yes. Though some educators may argue otherwise, the ambiguity of even fully vocalized Arabic and the need to convey stress and neutralization indicate that romanization is indeed necessary.

There is no question that learners must learn to read the Arabic script, eventually without the help of romanization or vowel diacritics. There is also no question that even professional Arabic teachers often wrongly assume that vocalized Arabic is *sufficient* for pronouncing Arabic correctly because vowel diacritics indicate the short vowels and double consonants. This is a common but unfortunate misconception based on the unconscious knowledge that native speakers have of the relation between the orthography and the phonology. To pronounce Arabic correctly, one must pay careful attention to the following phonological, phonetic and prosodic features:

1. **Phonemes:** how to pronounce each phoneme correctly.
2. **Words stress:** which syllable to stress.
3. **Neutralization:** which long vowels to shorten and which double consonants to undouble.
4. **Allophones:** select the correct allophone for the context in question.

The Arabic script is not capable of representing the last three features at all, and even fully vocalized Arabic cannot unambiguously represent the first. For these reasons, a romanized transcription is highly desirable to facilitate the learner in the critical initial stages. However, since none of the existing phonemic transcription systems provide sufficient detail to do this effectively, there is a need for a new system that addresses these shortcomings.

2. Description of System

A innovative romanization system, referred to as the **CJKI Arabic Romanization System**, or ***CARS*** for short, was developed and tested by The CJK Dictionary Institute (CJKI) specifically for the ease of use by learners. Since CARS provides precise phonemic transcriptions, it is also suitable for analyzing the phonological structure of words. The new system, which is making its debut with the appearance of *The CJKI Arabic Learner’s*

Dictionary (Halpern 2009c), has some unique features not found elsewhere, including a user-friendly set of symbols that are both easy to read and represent Arabic phonemes accurately. For a quick glimpse of how CARS works, study the following phrase:

اَلْحُكُومَةُ اَلْيَابَانِيَّةُ *ʾalḥukū́matu‿lya̱ba̱níyyatu* the Japanese government

Long vowels, as *ū́* in اَلْحُكُومَةُ (*ʾalḥukū́matu*), are shown by a macron over the vowel. Word stress is indicated by the accent mark, as in *níy* and *kū́* in *ʾalḥukū́matu‿lya̱ba̱níyyatu* above. For the first time in any system, neutralized long vowels are shown by a macron below the vowel (as in *a̱* above) to remind the reader that they are written, but *not* pronounced, as long vowels. The undertie (‿) is used as a liaison before words beginning with *ʾalif waṣla* (ٱ)**.**

CARS is a state-of-the-art transcription system that aims to be easy to learn. CARS is both innovative and conventional; innovative in that it uses a small set of diacritics and special symbols to unambiguously represent correct pronunciation, word stress and neutralization; conventional in that on the whole it maintains compatibility with conventional systems so that the user has few new symbols to learn. CARS comes in three flavors:

1. The **Standard CARS** symbol set aims to (1) unambiguously represent the Arabic phonemes (*phonemicization*)**,** (2) accurately indicate word stress (*accentuation*)**,** and (3) explicitly indicate which vowels are shortened and which consonants are undoubled (*neutralization*).
2. **Extended CARS** adds three useful optional features to Standard CARS: (1) indication of case endings (*declension*), (2) indication of velarized /a/ (*velarization*), and (3) indication of syllable boundaries (*syllabification*).
3. **Proxy CARS** is a set of substitute symbols to facilitate input from a standard ASCII keyboard. These **proxy symbols** makes it easy to input, search and manipulate the symbols used in Standard CARS and Extended CARS.

CARS is primarily a phonemic transcription system whose principal goal is to represent the phonemes of Arabic as they occur in actual pronunciation. Except for the optional velarized /a/, which is important for learners to be aware of, and vowel neutralization, it does not attempt to represent phonetic variations or allophones.

3. Standard CARS

3.1 Standard CARS Symbols

The CARS character set consists mostly of lowercase letters of the Latin alphabet supplemented and certain auxiliary symbols and diacritics. These auxiliary signs are referred to as **CARS symbols.** Table 1 describes the Standard CARS symbols, as opposed to the letters of the alphabet. Note that only lowercase letters are used; that is, capitals are not used to distinguish letters, as is done in some systems. The table also shows **proxy symbols** (in parentheses as necessary) that can used as substitutes for CARS symbols.

Table 1. Standard CARS Symbols

Symbol	Unicode	Symbol Name	Proxy Symbol	Description
¯	U+0304	macron	*vv*	The **macron,** as in *ā ī ū* (*aa ii uu),* indicates a long vowel, as in سَافَرْتُ *sā́fártu*. A **double vowel** can be used as a proxy symbol, as in *saafa/rtu.*
´	U+0301	acute accent	/	The **acute accent**, as in *á í ú,* (a/ i/ u/) represents a stressed syllable, as in أَنَا *ʾána̱***.** A **slash** can be used as a proxy, as in *'a/na_.*
¯́	U+0304 + U+0301	macron + acute accent	*vv/*	The **macron plus acute accent,** as in *ā́ ī́ ū́* (*aa/ ii/ uu/),* indicates a long stressed vowel, as in اَلصَّرَّافُ *ʾaṣṣarrā́fu.* A **double vowel plus slash** can used as a proxy, as in *ʾaSSarraa/fu.*
ˍ	U+0331	macron below	*x_*	The **macron below,** as in *a̱ i̱ u̱ ḇ ḏ etc.* (*a_ i_ u_ b_ d_* etc.), represents neutralization: either a long vowel that is shortened, as in أَنَا *ána̱*, or a double consonant that is undoubled, as in حُبّ *ḥuḇ* (*Hub_)*. A **vowel plus macron** can used as a proxy, as in *'a/na_.*
.		underdot	*caps*	The **underdot** modifies certain consonants to represent others, especially the emphatics: i.e., ص *ṣ,* ض *ḍ,* ط *ṭ,* ظ *ẓ* and ح *ḥ.* **Capitals** can be used as a proxy, as in *S, D, T, Z, H.*
‿	U+203F	undertie	~	The **undertie** indicates liaison between words resulting from the omission of *ʾalif waṣla* (ٱ) in word initial position, as in فِي ٱلصِّينِ *fi̱‿ṣṣī́ni.* A **tilde** can be used as a proxy , as in *fi_~SSii/ni.*
ʾ	U+02BE		*none*	The **right half ring** represents the *hamza* in all its written forms (آ ئ ء ؤ أ إ) Thus أُ and أَ, إِ are represented by *ʾa, ʾi* and *ʾu,* as in أَخَذَ *ʾákhadha* (*a/khadha*).
ɛ	U+0025	epsilon	E	The Greek letter **epsilon** represents the letter *ɛayin* (ع), as in سَعِيد *saɛīd.* A **capital E** can used as a proxy, as in *saEiid.*
˘	U+02D8	breve	^	The **breve** disjoins the letters *t, s, d* and *k* from a following *h* to show that they are distinct letters when not separated by a vowel; that is, *t˘h* , *s˘h, d˘h* and *k˘h* represent ده, سه, ته and كه**,** as opposed to *th, sh, dh* and *kh,* which represent ث *thāʾ*, س *sīn,* ذ *dhāl,* and خ *khāʾ*. A **circumflex** can be used as proxy, as in *t^h s^h d^h k^h.*

3.2 Vowels and Diphthongs

The CARS symbol set used to represent vowels is conventional, rather than innovative, and follows the normal practice in phonemic systems for romanizing Arabic. The only innovation is the use of the optional symbol ***ɒ*** to represent a velarized /a/ (see 4.3).

Short vowels are simply represented by *a, i* and *u,* as in كُتِبَ *kútiba* ‘was written’. The two diphthongs are written as *ay* and *aw*, as in بَيْت *bayt* ‘house’ and يَوْم *yawm* ‘day’, while the two glides are written as *iy* and *uw,* as in عَرَبِيَّة *ɛrabíyya* ‘Arab’ and فُتُوَّة *futúwwa* ‘youth’. Long vowels are represented by placing a macron over the vowel, as for example in اَلصَّرَّافُ *ʾaṣṣarrā́fu* ‘money changer’ and بُيُوتٌ *buyū́tun* ‘houses’, rather than by doubling the vowel (e.g. *buyuut*). Neutralized vowels are represented by a macron below the letter, as in اَلْيَابَانُ *ʾalya̱bā́nu* ‘Japan’ (see 3.9), while stressed vowels are represented by the accent mark, as in كَتَبَ *kátaba ‘he wrote’* (see 3.8).

Nunation, *tāʾ marbūṭa* (ة), *ʾalif maqṣūra* (ى) and various other special vowels are represented strictly as they are pronounced, as in بُيُوتٌ *buyū́tun,* عَرَبِيَّة *ɛrabíyya* and عَلَى *ɛála̱,* with no attempt to indicate the original orthography. That is as it should be as CARS (with minor exceptions) is a phonemic, not graphemic, transcription system.

Table 2. Short Vowels

CARS	Proxy	Description
a	a	short /a/
i	i	short /i/
u	u	short /u/
á	a/	stressed short /a/
í	i/	stressed short /i/
ú	u/	stressed short /u/

Table 3. Long Vowels

CARS	Proxy	Description
ā	aa	long /a/
ī	ii	long /i/
ū	uu	long /u/
ā́	aa/	stressed long /a/
ī́	ii/	stressed long /i/
ū́	uu/	stressed long /u/

3.3 Consonants

Certain minor innovations were made in representing the consonants, including the use of *ɛ* for the letter ع and representing *hamza* phonemically in all contexts. However, to maintain compatibility with conventional romanization systems, on the whole the commonly used Latin letters used to transcribe Arabic consonants were also adopted in CARS, as shown below.

Table 4. Plain Consonants

ب	b	bāʾ
ت	t	tāʾ
ج	j	jīm
د	d	dāl
ر	r	rāʾ
ز	z	zāy

س	s	sīn
ع	ɛ	ɛayin
ف	f	fāʾ
ق	q	qāf
ك	k	kāf
ل	l	lām

م	m	mīm
ن	n	nūn
ه	h	hāʾ
و	w	wāw
ي	y	yāʾ
ء	ʾ	hamza

The underdot is used to modify the base consonant, especially the emphatic/pharyngealized consonant series, while digraphs are used to represent the consonants shown below.

Table 5. Underdotted Letters and Diagraphs

ص	ṣ	ṣād
ض	ḍ	ḍād
ط	ṭ	ṭāʾ
ظ	ẓ	ẓāʾ
ح	ḥ	ḥāʾ

ث	th	thāʾ
خ	kh	khāʾ
ذ	dh	dhāl
ش	sh	shīn
غ	gh	ghayn

Shadda is represented by duplicating the consonant that bears the *shadda,* as in مُحَمَّد *muḥammad*, written with a double *m.*

3.4 Consonant Assimilation

Some romanization systems do not indicate article assimilation explicitly; that is, the article اَلْ *ʾal* is spelled *al*, as in أَلصَّغِير *al-saghiir.* Since CARS is a phonemic system, when the article اَلْ precedes a sun letter the assimilated pronunciation is shown explicitly by doubling the letter in question, as in أَلصَّغِير *ʾaṣṣaghīŕ.*

3.5 Hamza

The **right half ring** represents the *hamza* (the glottal stop), rather than the conventional apostrophe. Thus ,أَ أُ, إِ, and أْ (and their variants آ ئ ء ؤ) are represented by *ʾa, ʾi* and *ʾu,* and ʾ*,* as in أَخَذَ *ʾákhadha.* Unlike other romanization systems that often omit the *hamza* in word initial position, CARS retains it in all contexts to remind the reader that it is a genuine consonant that must always be pronounced.

The problem with the apostrophe is that it is often omitted and that it is not sufficiently prominent. It is also easily confused with the various kinds of apostrophes used to transcribe *εayin.* We considered representing it with a letter, such as the Greek letter delta Δ or lambda Λ, but felt that this would be too radical a departure from convention and would be confusing since أَ and أُ would be spelled as *Δa* and *Δu* or as *Λa* and *Λu,* probably not very intuitive. For now we have settled for right hald ring. The symbolization of hamza requires further research. Suggestions for a new symbol are welcome.

3.6 Ɛayin

The letter ***εayin*** (ع) is probably one of the most difficult for non-Arabs to pronounce, and at the same time is also one of the most commonly used letters in Arabic. CARS uses the Greek letter epsilon ***ε*** to represent ع (as in سَعِيد *saεīd*), to which it is quite similar in shape, instead of the often confusing different types of apostrophes or such symbols as *3*. This is a welcome feature that ensures that ***εayin*** is pronounced at all times.

3.7 Digraph Disjunction

A diagraph is a sequence of two letters used to transcribe a single phoneme, such as ***sh*** for ش ***shiin*** and ***th*** for ث ***thā'.*** The problem with diagraphs is that sequences like ***sh*** are ambiguous. That is, ***sh*** can stand for ش ***shiin,*** or for the rarer combination of س ***sīn*** + ه ***hāʾ.*** For example, the word أَسْهَل 'easier', if it were romanized as *ʾáshal,* would seem as if it should be pronounced *ʾá·shal* rather than *ʾás·hal.* To disambiguate digraphs, some systems use an underline, as in sh for ش, to distinguish ***shīn*** from ***sīn*** + ***hāʾ***(*sh*)***;*** that is, to show that the two letter constitute a single orthographic unit (diagraph). But this creates a lot of noise since ***shīn*** is much more common than ***sīn*** + ***hāʾ*** . CARS takes the opposite approach: it uses the breve symbol (˘) to *disjoin*, rather than to join, the letters *t, s, d* and *k* from a following *h,* thus indicating that they are distinct letters; that is, ***t˘h***, s***˘h, d˘h*** and ***k˘h*** represent ده, شه, ته, and كه**,** as opposed to *th, sh, dh* and *kh,* which represent ث ***thāʾ***, س ***sīn,*** ذ ***dhāl*** and خ ***khāʾ***.

3.8 Word Stress

A major innovation of the CARS system is the indication of which syllable is stressed by placing an acute accent over the nuclear vowel, as in يَعْمَلُ ***yáεmalu*** 'he works'. Arabic stress rules are of great academic and theoretical interest, and have practical applications in pedagogy, speech technology and lexicography. However, almost all grammar books give stress rules that are inadequate or incomplete, giving the erroneous impression that stress can be easily predicted (Halpern, 2009b).

Though in many cases, such as كَتَبَ ***kátaba*** 'he wrote', stress is easy to determine from the rules, ultimately there is no way to determine stress unless one knows a set of complex rules and exceptions. An example of this is the shift in stress resulting from nunation with ةٌ ***tun***. If the case ending is pronounced, as in مَكْتَبَةٌ 'library' ***mak·tá·ba·tun*** (the middle dot indicates syllable boundaries), the antepenultimate syllable ***tá*** is stressed. But if the case ending is omitted, as is normal in speech, this word is pronounced ***mák·ta·ba,*** so that the stress shifts backwards from ***ta*** to ***mak*** (except in the Egyptian accent of MSA).

To avoid these complexities, CARS makes the stress explicit by using the acute accent. Though word stress is a major feature of CARS, for the sake of simplicity the accent can be omitted when it is not necessary, just like is done for the optional features of Extended CARS.

3.9 Neutralization

Another major innovation of CARS is the indication of which vowels and consonants are *neutralized* (shortened) in actual pronunciation by placing a macron below the neutralized letter, as in هٰذَا *hā́dha̱* ‘this’ (long final vowels are normally neutralized). This macron signifies that the vowel in question is potentially long, but is realized as short in this particular context.

Table 6. Neutralized Vowels

CARS	Proxy	Description
a̱	a_	neutralized /a/
i̱	i_	neutralized /i/
u̱	u_	neutralized /u/
ɒ̱	A_	neutralized velarized /a/

Double consonants at the end of a word are also neutralized, a phenomenon almost never discussed in the literature (Holes, 2004). Thus a word like حُبّ ‘love’, which is theoretically pronounced *ḥubb*, is in fact pronounced *ḥub,* as if it had no *shadda.* In CARS, this is represented by *ḥuḇ,* with the macron below the *b* indicating that the double consonant is realized as a single consonant (of course if the case endings are pronounced these consonants retain their full double value, i.e. حُبٌّ *ḥubbun).*

Neutralization is a very common phenomenon. Though Arabic teachers in carefully enunciated speech may sometimes pronounce neutralized vowels and consonants as long, this is an act of hypercorrection that does not occur in their own natural speech. Ignoring neutralization, as is done in all dictionaries and romanization systems, results in “theoretically correct” but unnatural or stilted pronunciation. For example, أَنَا *ʾána̱* ‘I’ and هٰذَا ***hā́dha̱*** ‘this’ are never pronounced with a final long /a/ (*ā),* while the first long vowel of اَلْيَابَانُ *ʾalya̱bā́nu* 'Japan' is shortened in pronunciation, shown by *a̱.* But dictionaries and textbooks transcribe these vowels as long, misleading one to believe that that is how they should be pronounced.

Neutralization is a complex phenomenon with complex rules that vary somewhat with the context and with the speaker, and includes such subtle phenomena as *half long* vowels (Halpern, 2009b). Though neutralization is a major feature of CARS, for the sake of simplicity the macron below can omitted when it is not necessary, just like is done for the optional features of Extended CARS.

3.10 Liaison

The undertie symbol, as in *a‿b,* indicates assimilation between words in liaison that occurs when the second word begins with an *ʾalif waṣla* (ٱ). The undertie replaces the *ʾalif waṣla,* which is not pronounced. For example, in فِي ٱلصِّينِ *fi̱‿ṣṣī́ni* ‘in China’, it shows that ٱ is not pronounced so that the two words are joined into a single phonological unit (in this case the ل *lām* is also assimilated too because it is a sun letter). This causes the syllable boundary to span across words, i.e., *fi̱‿ṣṣī́ni* is pronounced as three syllables: *fi̱‿ṣ·ṣī́·ni*.

4. CARS Extensions and Tools

4.1 Extended CARS

The functionality of CARS is enhanced by several optional features referred to as **Extended CARS.** This consists of three symbols to explicitly indicate (1) **declension** (case endings), (2) **vowel velarization** (velarized /a/), and (3) **syllabification** (syllable boundaries). These features are not normative (not part of the core CARS specification), but are useful for pedagogical and lexicographic applications as well as for phonological analysis. Extended CARS makes use of the following symbols.

Table 7: Extended CARS Symbols

Symbol	Unicode	Symbol Name	Proxy Symbol	Description
-	U+002D	hyphen	-	The **hyphen** is an optional symbol that indicates case endings, as in اَلصَّرَّافُ *ʾaṣṣarrā́f-u,* where the "-*u*" indicates the beginning of a nominative case ending. Parentheses can be used as proxy symbols, as in *ʾASSArrA/f(u).*
ɒ	U+0252	turned alpha	A	The **turned alpha** is an optional symbol that represents a velarized /a/, as in اَلصَّرَّافُ *ʾɒṣṣɒrrɒ̄́fu.* This vowel can be short (*ɒ*), long (*ɒ̄*), stressed (*ɒ́*), neutralized (*ɒ̱*) or both long and stressed (*ɒ̄́*). A **capital A** can be used as a proxy, as in *ʾASSArrAA/fu.*
·	U+00B7	middle dot	+	The **middle dot** is an optional symbol that indicates syllabification, as in اَلصَّرَّافُ *ʾaṣ·ṣar·rā́·fu,* used on special occasions to indicate syllable boundaries. A **plus sign** can be used as a proxy, as in *ʻA+-Sar+rAA/+fu.*

Normally, it is not necessary to use the options provided by Extended CARS, but they can be used freely as necessary. Using them all together, as in

اَلصَّرَّافُ ٱلْآلِـيُّ *ʾɒṣ·ṣɒr·rɒ̄́·fu͜l·ʾā·líy·y-u* ATM

makes a CARS string look a bit intimidating because of the various symbols. In Standard CARS, this is written *ʾaṣṣarrā́fu͜lʾālíyyu,* which is more friendly. Rarely should it be necessary to use all three options together, though it does have the advantage of representing phonemic, phonetic, prosodic, syllabic and grammatical information.

4.2 Declension

The use of the **hyphen** is an optional feature to indicate case endings, as in اَلصَّرَّافُ *ʾaṣṣarrā́f-u* ‘the money changer’**,** where the “*-u*” indicates the optional nominative case ending. In fully vocalized Arabic cases endings, including nunation, are shown, and are actually pronounced in highly formal speech. However, they are omitted in normal spoken MSA. Pronouncing case endings all the time sounds stilted and unnatural, and learners are advised not to do so. Thus أَخَذَ ٱلْقِرْدَ ٱلصَّغِيرَ *ʾákhadha‿lqírda‿ṣṣaghī́r-a* ‘he took the little monkey’ is actually pronounced as *ʾákhadha‿lqírda‿ṣṣaghī́r* (note that the *a* of *qírda* is not omitted because of article assimilation).

4.3 Vowel velarization

The use of the **turned alpha** (ɒ) is an optional feature to indicate **vowel velarization**. This refers to a low or central or back vowel, similar to the vowels in the English words *car* or *saw*. This **velarized /a/,** an allophone of /a/, can be short as in اَلصَّرَّافُ *ʾɒṣṣɒrrɒ̄́fu*, ‘money changer’, stressed (ɒ́), neutralized (ɒ̱) or both long and stressed as in صَارًا *ṣɒ̄́rɒ̱* ‘bacame’, usually occurs before or after the emphatic (velarized) consonants ط ظ ص ض)) and ق, sometimes after ر and rarely ل.

Table 8. Velarized Vowels

CARS	Proxy	Description
ɒ	A	velarized /a/
ɒ̄́	AA/	stressed long velarized /a/
ɒ̄	AA	long velarized /a/
ɒ́	A/	stressed velarized /a/
ɒ̱	A_	neutralized velarized /a/

4.4 Syllabification

The use of the **middle dot** to indicate syllabification, as in اَلصَّرَّافُ *aṣ·ṣar·rā́·fu,* is an optional feature useful for special occasions such as in dictionary headwords and linguistic discussions on word stress or syllabification. This feature should be used sparingly as it makes the transcription more difficult to read.

4.5 Proxy CARS

The most important feature of CARS is its unambiguous encoding of phonological and prosodic information in an intuitive, easy to read set of letters and symbols. However, this readability comes at a price: there are ten non-ASCII symbols that may require special effort to input from a standard keyboard. Some, like the acute accent to show stress, are commonly used in many languages, though inputting them can still be inconvenient; others, like the macron, the underdot and the undertie, can be challenging to input.

To this end, CJKI has developed a set of symbols consisting of pure ASCII characters that can be used as substitutes for genuine CARS symbols. These are referred to as **proxy symbols,** and CARS used in this way is called **Proxy CARS**. For example, inputting صغير ***ṣaghī́r*** in Standard CARS is not easy as it requires typing an *s* with an underdot and an accented *i* with a macron on top. This can be easily written in proxy symbols as *Saghii/r*. The proxy symbols are fully described in tables 1 and 7. They are summarized below with examples.

Table 9. Proxy Symbols

Symbol Example	Symbol Name	Proxy Example	Proxy Name	Function
ā	macron	*aa*	double vowel	long vowels
á	acute accent	*a/*	slash	stressed vowels
ā́	macron + acute accent	*aa/*	double vowel + slash	long stressed vowels
a̱	macron below	*a_*	underscore	neutralization
ṣ	underdot	*S*	capital letter	emphatics etc.
a‿b	undertie	*a~b*	tilde	liaison
d˘h	breve	*d^h*	circumflex	digraph disjunction
ɒ	turned alpha	*A*	capital *A*	velarization
ɛ	epsilon	*E*	capital *E*	*ɛayin*
la·ka	middle dot	*la+ka*	hyphen	syllabification

Proxy CARS can be used in three ways:

- 📂✍ As an input string in a CARS conversion software to convert proxy symbols to CARS symbols.
- 📄✍ When inconvenient to input CARS symbols, using the proxy symbols directly.
- 📃✍ For typing difficult to input symbols. That is, CARS symbols and proxy symbols can mixed in the same text, though this is undesirable.

Here is an example of Proxy CARS in action:

Arabic:	أَخَذَ عَلَاءُ ٱلدِّينِ ٱلْقِرْدَ ٱلصَّغِيرَ
English:	Aladdin took the little monkey
Standard CARS:	*ʾákhadha ɛalā́ʾu͜ ddī́ni͜ lqírda͜ ṣṣaghī́ra*
Proxy CARS:	*'a/khadha Ealaa/’u~ddii/ni~lqi/rda~SSaghii/ra*
Extended CARS:	*ʾá·kha·dha ɛa·lā́·ʾu͜ d·dī́·n-͜ l·qír·d-͜ ṣ·ṣɒ·ghī́·r-a*

As can be seen, Proxy CARS is not as easy to read as Standard CARS, but it certainly is much easier to type. However, both convey exactly the same phonological and prosodic information. Extended CARS conveys more information but is harder to read. In conclusion, Proxy CARS is easy to write and hard to read, whereas Standard CARS is easy to read and hard to write.

4.6 CARS Tools

CJKI has or is in the process of developing tools to automatically convert Proxy CARS to Standard CARS and Extended CARS, as well as tools to automatically convert from Arabic script or in Buckwalter transliteration to any of the three varieties of CARS.

1. **CARS IME**: This is an input method editor that converts proxy symbols to CARS symbols in real time, e.g. *ʾa/khadha Ealaa/ʾu~ddii/ni* is converted to *ʾákhadha ɛalā́ʾu͜ ddī́ni.*
2. **CARS Converter:** This batch conversion utility accepts text files written in proxy symbols as input and converts them to CARS symbols encoded in UTF-8. It can also convert CARS symbols to proxy symbols.
3. **CARS Macros:** Macros for Microsoft Word and Excel, as well as for Open Office Writer and Calc, can convert from CARS symbols to proxy symbols and vice versa.
4. **A2C Converter**: This accepts fully vocalized Arabic text files written in Arabic script and converts them into CARS symbols or proxy symbols. This powerful tool automatically determines the stressed and neutralized syllables, adds syllable and case ending boundaries, and marks velarized /a/.

Needless to say, the many phoneme-to-grapheme ambiguities mean that there is no way to convert CARS *to* Arabic script since CARS is a set of phonemic symbols that do not map to the original Arabic graphemes. In a case like *maktába* (مَكْتَبَة), for example, there is no way to determine that the final *a* corresponds to *tāʾ marbūṭa* (ـَة) rather than to *fatḥa* (بَ).

5. Sample Text

Below is a sample of a text written in fully vocalized Arabic followed by Standard CARS, Extended CARS and Proxy CARS, and for reference the widely used Intelligence Community Standard romanization also know as IC (NGA *undated*), the English translation and the Buckwalter transliteration (Buckwalter *undated*), in common use in NLP.

Vocalized Arabic

مُنْذُ مِئَاتِ ٱلسِّنِينَ، كَانَتْ تَعِيشُ فِي ٱلصِّينِ أَرْمَلَةٌ فَقِيرَةٌ مَعَ ٱبْنِهَا ٱلصَّغِيرِ عَلَاءِ ٱلدِّينِ. كَانَ عَلَاءُ
ٱلدِّينِ يَبْلُغُ مِنَ ٱلْعُمْرِ اِثْنَيْ عَشَرَ عَاماً فَقَطْ، وَلَكِنَّهُ كَانَ يَعْمَلُ فِي مَحَلِّ خَيَّاطٍ لِيَكْتَسِبَ مَا يَعِيشُ
مِنْهُ هُوَ وَأُمَّهُ.

Standard CARS

múndhu miʾā́ti‿ssinī́na, kā́nat taɛī́shu fi̲‿ṣṣī́ni ʾarmálatun faqī́ratun máɛa‿bniha̲‿ṣṣaghī́ri ɛalā́ʾi‿ddī́ni. kā́na ɛalā́ʾu‿ddī́ni yáblughu mína‿lɛúmri ʾithnáy ɛáshara ɛā́man fáqaṭ, walakínnahu kā́na yáɛmalu fi̲ maḥálli khayyā́ṭin liyaktásiba ma̲ yaɛī́shu mínhu húwa waʾúmmahu.

Extended CARS (without syllabification)

múndhu miʾā́ti‿ssinī́n-a, kā́nat taɛī́shu fi̲‿ṣṣī́n-i ʾarmála-tun fɒqī́ra-tun máɛa‿bniha̲‿ṣṣɒghī́r-i ɛalā́ʾi‿ddī́n-i. kā́na ɛalā́ʾu‿ddī́n-i yáblughu mína‿lɛúmr-i ʾithnáy ɛáshara ɛā́m-an fɒ́qɒṭ, walakínnahu kā́na yáɛmalu fi̲ maḥáll-i khayyɒ̄́ṭ-in liyaktásiba ma̲ yaɛī́shu mínhu húwa waʾúmmahu.

Proxy CARS (corresponding to Standard CARS)

mu/ndhu mi’aa/ti~ssinii/na, kaa/nat taEii/shu fi_~SSii/ni ‘arma/latun faqii/ratun ma/Ea~bniha_~SSaghii/ri Ealaa/’i~ddii/ni. kaa/na Ealaa/’u~ddii/ni ya/blughu mi/na~lEu/mri ‘ithna/y Ea/shara Eaa/man fa/qaT, walaki/nnahu kaa/na ya/Emalu fi_ maHa/lli khayyaa/Tin liyakta/siba ma_ yaEii/shu mi/nhu hu/wa wa’u/mmahu.

IC Romanization

mundhu mi'ati al-sinina, kanat ta'ishu fi al-sini armala faqira ma'a bniha al-saghiri 'ala'i al-dini. kana 'ala'u al-dini yablughu mina al-'umri ithnay 'ashara 'aman faqat, walakinnahu kana ya'malu fi mahalli khayyat liyaktasiba ma ya'ishu minhu huwa wa'ummahu.

Buckwalter Transliteration

muno*u mi}aAti {ls~iniyna, kaAnato taEiy$u fiy {lS~iyni >aromalapN faqiyrapN maEa {bonihaA {lS~agiyri EalaA'i {ld~iyni. kaAna EalaA'u {ld~iyni yabolugu mina {loEumori Aivonayo Ea$ara EaAmAF faqaTo, walakin~ahu kaAna yaEomalu fiy maHal~i xay~aATK liyakotasiba maA yaEiy$u minohu huwa wa>um~ahu.

English

Hundreds of years ago, there lived in China a poor widow with her young son, Aladdin. Aladdin was only twelve years old, but he worked in a tailor's shop to support himself and his mother.

6. Conclusions

As we have seen, CARS is an innovative transcription system that enables the learner to pronounce Arabic correctly and enables the linguist to analyze the phonological and prosodic structure of words. In addition to several unique features including a set of easy-to-read symbols to represent Arabic phonemes unambiguously, for the first time word stress and neutralization are represented in an explicit manner. Though these two features are of great importance to pedagogy and lexicography, currently they are not indicated in reference materials such as dictionaries and textbooks. Equally important is the suite of CARS tools that can automatically predict stress and neutralization as well as generate CARS transcriptions from vocalized Arabic.

Arabic romanization is a topic of research that deserves the serious attention of linguists and educators. Our institute is contributing to this effort through the compilation of the world's first Arabic-English dictionary that indicates both stress and neutralization in the CARS transcription for each entry. Though the basic specifications of CARS have been completed, this new system is still somewhat of a work in progress. Specialists in Arabic information processing, linguists and educators are encouraged to collaborate by proposing modifications on how to further enhance the system.

References

AbdulJaleel, N.A. and Larkey, L. 2003. Statistical transliteration for English-Arabic Cross Language Information Retrieval. *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management*. New Orleans: ACM. 139-146.

Buckwalter, Tim, 2009, Buckwalter transliteration. Downloaded from *http://en.wikipedia.org/wiki/Buckwalter_Transliteration* *circa* December 15, 2009.

Janssens, G., 1972. *Stress in Arabic and Word Structure in the Modern Arabic Dialects*. Belgium: Peeters.

Halpern, J. 2007. The Challenges and Pitfalls of Arabic Romanization and Arabization. *Proceedings of the Second Workshop on Computational Approaches to Arabic Script-Based Languages*. Palo Alto.

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

Halpern, J. 2009b. Word Stress and Vowel Neutralization in Modern Standard Arabic, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. Cairo: MEDAR.

Halpern, J. 2009c. The CJKI Arabic Learner's Dictionary. Downloaded from *http://www.kanji.org/kanji/dictionaries/cald/cald_overview.pdf* circa June 15, 2009,

Holes, C., 2004. *Modern Arabic Structures, Functions and Varieties*. Washington, D.C.: Georgetown University Press.

Mitchell, T.F., 1990. *Pronouncing Arabic*. Oxford, U.K.: Oxford University Press.

NGA, 2009. National Geospatial-Intelligence Agency. Proposal submission dowloaded from *http://www.acq.osd.mil/osbp/sbir/solicitations/sbir061/nga061.pdf* circa December 1, 2009.