Applying Smartphone Technology to Compile Innovative Arabic Learner's Dictionaries

Jack Halpern
Director
The CJK Dictionary Institute
Niiza, Japan
jack@cjki.org

***Abstract*—Arabic dictionaries suffer from several drawbacks that render them mostly inadequate for learners. The advent of mobile digital devices, especially the smartphone platform, has made it possible to significantly enhance the language learning experience in hitherto unavailable ways. This paper describes some of the methodology used in compiling two innovative Arabic learner's dictionaries fine-tuned to the special needs of learners that present abundant lexicographic information in a user friendly manner. The first is a printed dictionary and the second is a mobile application for the smartphone platform. Both use a phonemic transcription system that enables learners to pronounce Arabic accurately.**

Keywords-Arabic dictionary, pedagogical lexicography, smartphone

I. BACKGROUND AND AIMS

The lack of effective learner's dictionaries and other pedagogical tools puts learners of Arabic as a second language at a disadvantage compared with those of other major languages.

Existing dictionaries, often rooted in Classical Arabic, suffer from shortcomings that make them inadequate for learners. These include, among other things, user-unfriendly design, lack of romanization, inaccurate or out-of-date equivalents, and ordering by roots. Another obstacle for learners is the lack of accurate phonological information, especially word stress. For beginners who have no knowledge of roots and patterns, locating entries in traditional dictionaries is time-consuming and unreliable. Equally inconvenient is the lack of illustrative examples and the absence of part-of-speech labels.

In light of the acute surge in Arabic language studies in recent years, there is a growing need for a dictionary that meets the practical needs of the learner. This paper introduces a new Arabic learner's dictionary, referred to as CALD below. The primary goal of CALD is to enable learners to gain a firm understanding of the basic meanings and grammatical features of the core vocabulary of Modern Standard Arabic. From the outset, CALD was designed to meet the specific needs of beginners and intermediate learners by addressing the shortcomings of existing dictionaries. CALD is also being made available as an electronic dictionary application for the smartphone platform (iCALD).

A number of features distinguish CALD as a new type of pedagogical tool, including a learner-oriented romanization system that indicates word stress, a user-friendly design, accurate up-to-date equivalents, abundant illustrative examples, detailed grammatical information such as part of speech codes, alphabetical ordering, and *full form* search in the smartphone application. Everything from the selection and presentation of headwords and their meanings to the writing of examples and the entry layout was fine-tuned to create an effective learning aid that stimulates a desire to learn.

II. COMPILATION TECHNIQUES

A. Methodology

Briefly, CALD was compiled with the help of computational lexicography techniques combined with the latest advances in digital publishing technology. Various tools were used to perform extensive sanity checks to ensure data integrity, accuracy and consistency, and the work was independently reviewed by a team native speaker editors and lexicographers.

CALD is firmly committed to the descriptive approach. Whereas existing dictionaries often include obscure or archaic meanings occurring only in Classical Arabic, the word senses in CALD were newly created on the basis of actual occurrences, not merely on the authority of other dictionaries.

B. Selection Criteria

The lemmata have been selected on the basis of existing dictionaries and corpora-based frequency statistics. These include *A Frequency Dictionary of Arabic* [1], based on a 30-million-word corpus, the lexical database used for the ElixirFM project [2], K. Honda's excellent Arabic-Japanese learner's dictionary [3] and the Arabic Treebank [4]. The single senses were selected on the basis of occurrences in various sources supplemented by the subjective judgment of the editors.

The selection of entries and word senses for learner's dictionaries is not a mechanical process that can be based on raw frequency alone. The final decision was based on "usefulness to the learner," as judged by the editors. As a result, CALD includes up-to-date lemmata and senses ignored by many dictionaries, such as مُدَوَّنَةٌ *mudawwánatun*, 'blog' and اَلصَّرَّافُ ٱلْآلِـيّ *ʾaṣṣarrā́f-u_l-ʾālíyy-u* 'ATM'.

C. Alphabetical Ordering

A serious obstacle to learners is that locating entries in traditional dictionaries requires knowledge of roots, making the lookup process time-consuming and unreliable. Though a

أَكَلَ ʾá•ka•la VT [أ-ك-ل]
eat

لَمْ يَأْكُلْ شَيْئًا مُنْذُ صَبَاحِ ٱلْيَوْمِ *lam yá'kul sháy'-an múndhu ṣabáḥ-i_lyáwm-i* He has not eaten anything since this morning.

Form I - Z1 - 123	AP آكِلٌ *ʾákil-un*
PR يَأْكُلُ *yá'kulu*ᵘ	PP مَأْكُولٌ *ma'kúl-un*
IM كُلْ *kul*	VN أَكْلٌ *ʾákl-un*

بَيْتٌ báy•t-un NOUN [ب-ي-ت]
① house, home, residence

بَيْتٌ صَغِيرٌ *báyt-un ṣaghír-un* small house

② verse

بَيْتُ ٱلشِّعْرِ *báyt-u_shshíɛr-i* verse of poem

MS بَيْتٌ *báyt-un*	MP ① بُيُوتٌ *buyút-un*
FS —	② أَبْيَاتٌ *ʾabyát-un*

صَرَّافٌ ṣar•rấ•f-un NOUN [ص-ر-ف]
money changer, cashier

يَعْمَلُ ٱلصَّرَّافُ فِي ٱلْبَنْكِ *yáɛmalu_ṣṣarráf-u fí_lbánk-i* The money changer works at the bank.

MS صَرَّافٌ *ṣarráf-un*	MP صَرَّافُونَ *ṣarrafúna*
FS صَرَّافَةٌ *ṣarráfa-tun*	FP صَرَّافَاتٌ *ṣarrafát-un*

اَلصَّرَّافُ ٱلْآلِيُّ *ʾaṣṣarráf-u_l'ʾalíyy-u* ATM (automated teller machine)

قَرِيبٌ qa•rí•b-un ADJ [ق-ر-ب]
① near, close

قَرِيبٌ مِنْ بَيْتِي *qaríb-un min báyti* near my house

② recent, soon

فِي وَقْتٍ قَرِيبٍ *fí wáqt-in qaríb-in* in a short time

MS قَرِيبٌ *qaríb-un*	MP قَرِيبُونَ *qarìbúna*
FS قَرِيبَةٌ *qaríba-tun*	CM أَقْرَبُ *ʾáqrab-u*

وَقَعَ wá•qa•ɛa VI [و-ق-ع]
① fall down, drop, tumble

وَقَعَ مِنَ ٱلسُّلَّمِ *wáqaɛa mína_ssúllam-i* He fell off the ladder.

② take place, happen, occur

وَقَعَ حَادِثٌ فِي ٱلْمَصْنَعِ *wáqaɛa ḥádith-un fí_lmáṣnaɛ-i* An accident happened in the factory .

③ be located, be situated, lie

تَقَعُ ٱلْمَكْتَبَةُ وَسْطَ ٱلْجَامِعَةِ *táqaɛu_lmaktába-tu wásṭa_ljámíɛa-ti* The library is situated at the center of the university.

Form I - A2 - 150	AP وَاقِعٌ *wáqiɛ-un*
PR يَقَعُ *yáqaɛu*ᵘ	PP مَوْقُوعٌ *mawqúɛ-un*
IM قَعْ *qaɛ*	VN وُقُوعٌ *wuqúɛ-un*

وَقَعَ عَلَى *wáqaɛa ɛálą* VI come across, run into

وَقَعَ فِي حُبٍّ *wáqaɛa fí ḥúbb-in* VI fall in love

اَلْيَابَانُ ʾal•yą•bấ•n-u NOUN
Japan

سَافَرَتْ إِلَى ٱلْيَابَانِ *sấfarat ʾílą_lyąbấn-i* She travelled to Japan.

يَابَانِيٌّ yą•bą•níyy-un NISBA
Japanese ADJ

اَلْحُكُومَةُ ٱلْيَابَانِيَّةُ *ʾalḥukúma-tu_lyąbąníyya-tu* Japanese government

a Japanese NOUN

سَأَلَنِي يَابَانِيٌّ سُؤَالًا *sa'álanį yąbąníyy-un su'ấl-an* A Japanese asked me a question.

MS يَابَانِيٌّ *yąbąníyy-un*	MP يَابَانِيُّونَ *yąbąnìyúna*
FS يَابَانِيَّةٌ *yąbąníyya-tun*	FP يَابَانِيَّاتٌ *yąbąnìyát-un*

Figure 1. Page sample of CALD.

knowledge of roots is valuable, looking up by roots is inconvenient for beginners since they are unable to identify the root of a verb or to determine the basic form (Form I) from a derived form.

For example, to look up the verb اِنْتَظَرَ *'intáẓara* 'wait for' the user needs to know that it is Form VIII, derived from the basic form نَظَرَ *náẓara* 'look at', and be able to identify the root as نظـر *n-ẓ-r*. Under that root, اِنْتَظَرَ is listed along with other verbs like أَنْظَرَ *'ánẓara* 'grant' (Form IV) and تَنَاظَرَ *tanáẓara* 'face each other' (Form VI). Looking for اِنْتَظَرَ under نظر is not intuitive and means that beginners, even after repeated time-consuming searches, may fail to find the desired entry.

An important feature of CALD is the ordering of entries in alphabetical order, so that اِنْتَظَرَ is found under its canonical form اِنْتَظَرَ rather than under its root نظـر. Many leading Arabists, such as the team of experts that set the editorial policy for Sharoni's comprehensive Arabic-Hebrew dictionary [5], are in favor of ordering alphabetically because it enables the general user (not just the learner) to locate entries quickly and easily. In CALD, as in Sharoni's dictionary, it is also possible to search by roots since cross-references point from all verb roots to the verbs derived from that root.

III. WORD SENSES

A. *Up-to-date Meanings*

Existing dictionaries often list meanings that are out-of-date, mixing historical meanings with modern ones with no indication of their temporal status. On the other hand, recent words and word senses, especially those related to information technology, are often omitted. Another issue is that in some dictionaries historical or arbitrary sense ordering makes it difficult to know which senses are important or current.

In CALD, the senses are listed in order of importance in contemporary Arabic, and the English equivalents are up-to-date, accurate and concise. When a new concept first enters a language, the meaning of an existing word is often extended to cover that of the new concept, as in سَاحَة *sáḥa*, 'courtyard, field', extended to mean 'forum', while in other cases loanwords like تِلِفُون *tilifún* 'telephone' are adopted. The editors of CALD have made an effort to include such recent senses, which are often not listed in other dictionaries.

For the convenience of the learner, the equivalent indicates the particles that collocate with intransative verbs, such as سَلَّمَ عَلَى *sállama εalạ* 'greet (someone)' and سَافَرَ إِلَى *sáfara ʾilạ* 'travel to', while subentries list useful word combinations such as phrasal verbs like قَامَ بِ *qáma bi*, 'take upon' and وَقَعَ عَلَى *wáqaεa εálạ* 'come across' and prepositional phrases like مِنَ ٱلْمُمْكِنِ *mina lmúmkini* 'possible'.

B. *Illustrative Examples*

A survey [6] showed that only three out of 30 bilingual Arabic dictionaries give example sentences for all or almost all entries (some of which include snippets from the media too difficult for learners), while a couple of dictionaries give examples only very occasionally. From a pedagogical point of view, quoting examples from corpora or directly from sources like newspapers can be counterproductive since such sentences can be long, complicated, or difficult to understand.

In CALD, each sense is illustrated by easy example sentences *written specifically for this dictionary*, from a pedagogical point of view. As can be seen from Figure 1, the examples enable the learner to gain a good understanding of how the headword is used in context.

IV. GRAMMAR AND MORPHOLOGY

A. *Grammatical Information*

CALD provides detailed grammatical information such as the regular and irregular inflected forms for the imperfect, active and passive participles, verbal noun(s), feminine form, sound and broken plurals, the root, and the like.

Form I - Z1 - 123	AP آكِلٌ *ʾákil-un*
PR يَأْكُلُ *yá'kulu*^{u}	PP مَأْكُولٌ *ma'kúl-un*
IM كُلْ *kul*	VN أَكْلٌ *ʾákl-un*

Figure 2. Grammar Box for أكَلَ

Multiple plurals corresponding to different word senses are also shown, e.g., the plural of بَيْتٌ *báytun* for the sense of 'house' is بُيُوتٌ *buyūtun* but أَبْيَاتٌ *ʾabyắtun* for the sense of 'verse'.

MS بَيْتٌ *báyt-un*	MP ① بُيُوتٌ *buyắt-un*
FS —	② أَبْيَاتٌ *ʾabyắt-un*

Figure 3. Grammar Box for بَيْتٌ

The Grammar Box thus provides the learner with detailed information on inflected and derived forms, eliminating the need for laboriously consulting grammar books and conjugation tables.

B. *Part of Speech Codes*

The previously mentioned survey [6] showed (1) that 22 out of 30 dictionaries do not provide explicit POS codes, including such well-known dictionaries as Hans-Wehr [7], and (2) that not a single dictionary gives explicit verb transitivity codes for all verb entries (although a couple provide them implicitly). The failure of many dictionaries to provide POS codes is linguistically untenable and inconvenient to learners.

An important feature of CALD is that explicit and accurate part of speech codes are given for every entry. For example, many *nisba* adjectives have noun counterparts, such as يَابَانِيٌّ *yạbạnīyyun* 'Japanese' meaning 'a native of Japan' as a noun and 'relating to Japan' as an adjective. These are listed separately, with separate example sentences, to show how each is used in context. A useful feature not found elsewhere is that CALD indicates verb transitivity by explicit part-of-speech codes such as *VT* and *VI*.

V. PHONOLOGY AND ROMANIZATION

There is a misconception that vocalized Arabic is sufficient for pronouncing Arabic correctly and therefore learners should avoid romanization. However, Halpern [8] has shown that even fully vocalized Arabic cannot convey pronunciation accurately, and that a romanized transcription is highly desirable in the critical initial stages.

CALD introduces a new phonemic transcription system, called CARS, developed specifically to enable learners to pronounce Arabic accurately and with ease [8]. It offers several unique features, including a symbol set that unambiguously represents all Arabic phonemes, and, for the first time, symbols that explicitly indicate word stress and neutralization.

For example, in اَلْيَابَانِيَّةُ اَلْحُكُومَةُ *alḥukū́matu lyaḇaníyyatu* 'the Japanese government', long vowels (as in *kū*) are shown by a macron, word stress by the accent mark (as in *kú*), and, for the first time, long vowels that are *neutralized* (shortened) in actual pronunciation are indicated by an underline (as in *a̱*).

Though grammar books do give stress rules, these are inadequate and give the erroneous impression that stress is easily predictable [9]. To avoid these complexities, CARS makes stress explicit by placing an acute accent over the stressed syllable, as in يَعْمَلُ *yáɛmalu* 'he works'.

Neutralization has been neglected by dictionaries and learning materials. Learners are misled into believing that long vowels, as in أَنَا *ʾána* and كَتَبُوا *kátabu*, are pronounced long, and this myth is perpetuated by misguided educators who hypercorrect by pronouncing these vowels long in recordings.

A unique feature of CARS, which is of great utility to the learner, is the explicit indication of both stress and neutralization, as in هٰذَا *hắdha*. Ignoring neutralization, as is the common practice, results in "theoretically correct" but unnatural or stilted pronunciation. CARS encourages learners to pronounce correctly by making shortened vowels explicit.

Below is a sample of a text written in fully vocalized Arabic followed by the CARS transcription and the English translation.

كَانَ عَلَاءُ ٱلدِّينِ يَبْلُغُ مِنَ ٱلْعُمْرِ اِثْنَيْ عَشَرَ عَاماً فَقَطْ، وَلَكِنَّهُ كَانَ يَعْمَلُ فِي مَحَلِّ خَيَّاطٍ لِيَكْتَسِبَ مَا يَعِيشُ مِنْهُ هُوَ وَأُمَّهُ.

kắna ɛalāʾu_ddīńi yáblughu mína_lɛúmri ʾithnáy ɛáshara ɛắman fáqaṭ, walakínnahu kắna yáɛmalu fi̱ maḥálli khayyáṭin liyaktásiba mą yaɛīśhu mínhu húwa waʾúmmahu.

Aladdin was only twelve years old, but he worked in a tailor's shop to support himself and his mother.

As can be seen, the CARS transcription enables the reader to pronounce Arabic with precision, shortening vowels when necessary and stressing the correct syllable.

VI. SMARTPHONE TECHNOLOGY

CALD is also being made available as an electronic dictionary application for the smartphone (iOS and Android) platform, referred to as iCALD, which takes advantage of the superb features of the smartphone. A drawback of many existing dictionaries is poor design and user-unfriendliness. Often the fonts are so small and the design so poor that looking up words becomes a burdensome chore, strongly demotivating learners. Special efforts were made to design both CALD and iCALD for maximum user friendliness and portability. Typographical design with the aid of state-of-the-art digital publishing technology was used to achieve a harmonious blend of font styles and sizes, resulting in an esthetically pleasing design that strongly stimulates a desire to learn.

Space does not permit a description of iCALD in detail. Briefly, it offers a user-friendly interface, *full form* search for locating *any* inflected form, multiple search modes, audio for all lemmas and example sentences, and immediate access to an application that provides full conjugation paradigms.

VII. FUTURE WORK

This paper introduces Arabic learner's dictionaries that promise to transform the traditional methods of learning Arabic as a second language. These dictionaries satisfy the special needs of non-native learners by addressing the major shortcomings of previous works. Every effort has been made to meet those needs by providing such features as a learner-oriented romanization system, part of speech codes, alphabetical ordering, user-friendly design, illustrative examples, and a smartphone application with a user-friendly interface and full form search for locating any inflected form.

The number of learners of Arabic worldwide is steadily increasing, which has led to a constantly growing demand for pedagogically effective tools. Our institute is dedicated to continuing to meet this challenge through the ongoing development of learner's dictionaries and pedagogical applications. It is hoped that lexicographers and educators around the world will contribute to this effort through advice, constructive criticism and, above all, direct collaboration.

REFERENCES

[1] T. Buckwalter and D. Parkinson, A Frequency Dictionary of Arabic: Core Vocabulary for Learners, London: Routledge, 2011.

[2] O. Smrž, Functional Arabic Morphology: Formal System and Implementation, Doctoral Thesis, Prague: Institute of Formal and Applied Linguistics, Charles University in Prague, 2007.

[3] K. Honda, Pasupooto shokyuu arabiago jiten (Passport Beginner's Arabic Dictionary), Tokyo: Hakusuisha, 1997.

[4] M. Maamouri, A. Bies, T. Buckwalter, H. Jin, Treebank: Part 2 v 2.0. Linguistic Data Consortium, catalog number LDC2004T02, ISBN 1-58563-282-1, 2002.

[5] A. Sharoni, The Comprehensive Arabic-Hebrew Dictionary, Tel-Aviv: University of Tel-Aviv, 1987.

[6] The CJK Dictionary Institute, Inc. (CJKI). "Arabic Dictionaries". Internal survey. July 6, 2011.

[7] J. Cowan, Hans Wehr: A Dictionary of Modern Written Arabic, Beirut: Librairie du Liban, 1979.

[8] J. Halpern, CJKI Arabic Romanization System, Abu Dhabi: The International Symposium on Arabic Transliteration Standard: Challenges and Solutions, 2009.

[9] J. Halpern, Word Stress and Vowel Neutralization in Modern Standard Arabic, 2nd International Conference on Arabic Language Resources and Tools, Cairo: The MEDAR Consortium, 2009, pp. 42-47.