Headword Selection in Arabic Lexicography

by Jack Halpern
The CJK Dictionary Institute, Japan

This monograph discusses key issues related to the selection of headwords in Arabic dictionaries, in particular learner's dictionaries, and briefly touches on criteria for selecting word senses. In addition, these issues are discussed as they pertain to the brand new CJK Arabic-English Learner's Dictionary. Since our goal is to summarize major points, methodological details and formal citations have been omitted (these will be included in a forthcoming paper).

Traditional selection criteria

Headword selection is fundamental to the practice of lexicography. Particularly for Arabic lexicographers, who must at once deal with an overabundance of older words and an influx of newer words, determining the criteria for selection becomes a daunting task. This problem has only recently begun to be addressed using modern, corpus-based computational methods, an approach spearheaded by the *Oxford Arabic Dictionary* (OAD).

Broadly speaking, selection criteria fall into five principal categories:

1. Drawing from the Quran and classical literature
2. Copying from past dictionaries
3. Analyzing authentic texts found in corpora
4. Relying on the subjective judgment of the lexicographer
5. Determining "importance" based on both frequency and the lexicographer's discretion

The majority of traditional Arabic dictionaries, and many modern ones, have relied heavily on the first two of these methods. Headwords have been selected in large part from the Quran, poetry, and classical texts, and lexicographers have added newer words as they have seen fit.

While dictionaries compiled in this manner may certainly serve their purposes as aids in reading classical texts, they are poorly suited for those studying Modern Standard Arabic. As a rule, new words, especially those related to technology and modern life, have been conspicuous by their absence. When they are included at all, they are often buried among scores of archaisms.

Modern selection criteria

In order to compile a truly practical dictionary of Modern Standard Arabic, one clear solution is the third of the above selection criteria: that is, the use of corpora to choose entries based on authentic texts. This is the approach taken by the recently

published OAD, which serves admirably as a comprehensive reference for modern Arabic. This strategy finds its greatest advantage in its objectivity: lexicographers can rely on computational algorithms to automatically tokenize, select, and lemmatize on the basis of occurrence in corpora without excessively depending on the lexicographer's subjective judgment.

However, from the point of *pedagogical* lexicography, one of whose goals is to provide only the most useful headwords for a learner, fully comprehensive dictionaries are not ideal. Even superior works such as the OAD can prove to be overwhelming or confusing for learners. Instead, learners are best served by more practically-oriented criteria for headword selection.

To this end, some dictionaries choose to use a statistical approach, choosing words as dictated by their frequencies of occurrence in corpora. This is the strategy chosen by Buckwalter and Parkinson in their *A Frequency Dictionary of Arabic*, and once again, this is a major step in the right direction.

However, even this frequency-based approach can ultimately prove a double-edged sword for three reasons, discussed in more detail below: the omission of multiword expressions, the omission of derived forms, and the mistaken inclusion of infrequent words.

Multiword expressions

First, using a system of frequency of occurrence statistics based on *orthographic words* (a sequence of letters delimited by spaces) leads to the omission of critically important *multiword expressions* (MWEs). An MWE can be defined as two or more words that together function as a single lexical unit, for instance a compound word or idiomatic expression. While spaces are of course significant for delimitation, their presence or absence cannot independently determine the status of a word as a lexical unit. Nevertheless, the simplicity of orthographic words has established them as the basic units of analysis of Arabic corpora.

Many MWEs are, however, highly useful, and many are used as often as or more often than single orthographic words. For example, *high school* and *school bus* are frequent compound words, and each should be lemmatized and considered a single dictionary entry (rather than an example or subentry under *high* or *school*). Both of these are written as two orthographic words, while other, similarly constructed words like *headwaiter* have become single orthographic words purely by historical happenstance. There is therefore no reason for words like *headwaiter* to be prioritized as dictionary entries over MWEs like *high school* simply because they are written as single orthographic words.

Omission of multiword expressions

Listing MWEs as normal dictionary entries is standard practice among lexicographers of Western languages, and as a result, few Western language dictionaries equate orthographic words with dictionary headwords. This is, however, not the case in Arabic lexicography, where traditionally the root and canonical forms served as the basic units of lemmatization with little or no attention to MWEs.

Recent Arabic lexicography, which is focusing increasingly on corpus analysis, has been unable to avoid this problem. The process of tokenization – used for segmenting a stream of text in corpora on white spaces and punctuation marks – inevitably leads to the ignoring of MWEs. Compounds equivalent to *high school* are misleadingly processed as new instances of the separate words *high* and *school*. This is ultimately an inaccurate reflection of the language, as it both overlooks important MWEs and overcounts the occurrence of the individual components.

This problem is particularly pressing for Arabic lexicography, since MWEs are common in both modern and classical Arabic. For example, اَلشَّرْقُ ٱلْأَوْسَطُ *'assharqu l'awSaTu* 'the Middle East', a compound word, and اَلْحَمْدُ لِلّٰهِ *'alHamdu lillahi* 'Thank God', a common fixed expression, are both high frequency MWEs. However, such lexical units, if they are included in Arabic dictionaries at all, are given only as examples under one of their components (for instance under شَرْق 'east' or حَمْد 'praise').

Even works based on advanced corpus analysis such as the OAD have not systematically included MWEs as inclusion would require a new set of analytical tools not yet available, increasing both technical difficulty and cost. Particularly difficult to analyze are *discontinuous* MWEs, analogous to *took off* in the phrase *he took his jacket off* in English. For instance, in the phrase نَعْتَذِرُ مِنْكُمْ عَنْ *na`tadhiru minkum `an* 'We apologize for...', the particle عَنْ *`an* does not immediately follow the verb نَعْتَذِرُ *na`tadhiru.* It remains clear, however, that, whether as headwords or as subentries, the inclusion of common MWEs is in the best interest of both the general user and the language learner.

Omission of derived words

Yet another problem results from *stemming,* the process of reducing inflected or derived forms to their "stem" – generally the canonical form of a lexeme – and *decliticization,* or the removal of clitics such as prepositions. Stemming is of course a necessary step in dictionary compilation: one of the most basic tasks of the lexicographer is to identify that, in the case of English, the forms *eat*, *eating*, *eats*, and *eaten* belong to a common lexeme class, and *eat* may be selected as the lemma representing that class (the canonical form).

This presents no problem as long as stemming is limited to conjugated forms of verbs as well as inflected forms and declensions of nouns and adjectives. In Arabic lexicography, however, past dictionaries have generally fallen towards overly aggressive stemming and decliticization of *derived* words – in other words, unconditionally including only base forms and omitting forms derived from nouns. For instance, while all dictionaries include the word سُرْعَة *sur'a* 'speed', few dictionaries systematically include the major derived forms such as بِسُرْعَةٍ *bisur'atin*, literally 'with speed' but commonly used as an adverb meaning 'quickly'. Such derivations are useful to both learners and general users alike, since بِسُرْعَةٍ 'quickly' and سُرْعَة 'speed' are used at comparable frequencies.

Derived words thus have a place in both general and learner's dictionaries, while inflected and declined forms (e.g. case endings and plurals) may usually be safely omitted. Some dictionaries, including the OAD, do occasionally include derivations as examples under their respective lemmas, but once again, these are often included unsystematically at the whim of the editor. Moreover, learners searching for such derivations as *bisur'atin* in a paper dictionary can find the relevant example – located under the root or the canonical form – only if they have a strong grasp of Arabic grammar. While electronic dictionaries can automatically direct users to the appropriate entry, this is a fundamental flaw for paper dictionaries, which leave learners struggling to find common words.

Inclusion of infrequent words

Finally, almost as problematic as the omissions discussed above is the inclusion of words and senses that are rarely used in modern Arabic.

Arabic includes a vast number of words and countless synonyms accumulated and preserved over many centuries within the dusty tomes of classical dictionaries. Many Arabic scholars look to the 13th century dictionary *Lisān al-'Arab* as a standard reference, despite the fact that its contents are at times irrelevant to modern Arabic. Rather than deleting such archaic words as they have gone out of use, however, classical Arabic lexicographers have held on to them tenaciously. As a result, modern Arabic lexicographers now face the formidable task of separating the wheat from the chaff to ensure that the headwords selected reflect contemporary usage.

Clearly, up-to-date corpora should play a critical role in this effort, but utilizing corpora effectively is no easy task. A frequency-based approach that selects entries based on token counts alone leads to the inclusion of words that are used frequently as components of compounds but used seldom, if ever, on their own. One example is تِشْرِين *tishriin,* which is not used by itself but is a common component of such compound words as تِشْرِين الأَوَّل *tishriin 'al'awwal* 'October' and تِشسرينُ ٱلثَّانِي *tishriin-aththaani* 'November'.

Another example is the word حَمْد *Hamd* 'praise' in اَلْحَمْدُ لِلّٰهِ *'alHamdu lillaahi* 'Thank God'. Although this phrase is extremely frequent in modern Arabic, *Hamd* is much less common as an independent word. Thus a pure statistical approach based on counting tokens would misleadingly assign *Hamd* a high frequency simply because of its occurrence in the full phrase. Indeed, Buckwalter and Parkinson's *Frequency Dictionary* lists this word at a frequency of 296 out of 5000, misleadingly signaling to the learner that the word is a high priority in its own right. Lexicographers must thus be careful to avoid the trap of listing such components as independent dictionary entries simply because they occur frequently.

As a side note, the use of corpora does not – on the other side of the coin – guarantee the inclusion of all frequent words and expressions. Even Buckwalter and Parkinsons's excellent *Frequency Dictionary* is missing such common words as أَذِنَ *'adhina* 'to permit' while on the other hand it lists many rarer words not useful to learners . This once again highlights the fact that corpus statistics require careful editing and supplementation in order to be of most benefit to learners.

Pedagogical selection criteria

In compiling **The CJK Arabic-English Learner's Dictionary** (CALD), we have avoided the potential pitfalls above, including overreliance on frequency statistics and undue dependence on classical sources. Based on a systematic approach, this dictionary is designed to meet the specific needs of non-native learners of Arabic in the beginner to intermediate levels.

To achieve these aims, we have made a special effort to include frequent words, derivations and multiword expressions drawn from contemporary Arabic corpora, textbooks for learners, modern learner's dictionaries, the web, and other sources. To ensure that the selected entries and word senses truly reflect standard contemporary usage, and furthermore to ensure that they are truly useful to the learner, each entry and sense was checked and double-checked by a team of lexicographers experienced in Arabic pedagogy and deeply versed in Arabic grammar. This team validated the occurrence of entries and word senses as they occur *in the living language* while avoiding the common practice of including words and senses merely on the authority of other, often classical sources.

To summarize, we have brought CALD in line with modern pedagogical lexicography by following the guidelines below.

1. Potential headwords were selected through analysis of corpora and modern pedagogical materials such as learner's dictionaries, textbooks and the web.
2. Special efforts were made to include modern words such as *blog* and *ATM*.
3. Rare and archaic words were, for the most part, omitted by relying on modern corpora and sources.
4. Dialectical words were omitted regardless of frequency within the corpora.

5. Headwords and senses were further culled by a team of lexicographers experienced in Arabic pedagogy to maximize relevance to learners.
6. Multiword expressions were systematically included as main entries or subentries rather than as examples.
7. Word senses were listed in order of importance to accurately reflect modern usage.
8. Important derived words, especially adverbs, were systematically included.

The above approach has allowed us to enjoy the benefits of corpora-based lexicography – namely the omission of rare and archaic words – while avoiding dialectal words and high frequency orthographic words that are seldom used on their own. Furthermore, the systematic inclusion of important derived words, for instance adverbs of the pattern بِبُطْءٍ *bibut'in* 'slowly' and رَغْمًا *raghman* 'in spite' (not even found in comprehensive dictionaries) enables even users of paper dictionaries to quickly search for such words without knowledge of their roots.

Conclusion

Throughout the compiling process, we have striven to ensure that CALD is fully in keeping with the principles of pedagogical lexicography. We have made great efforts to create a dictionary in which learners can easily find the words most relevant to them, whether they be roots, multiple word expressions, or derived words. After all, dictionary users, and in particular learners, are generally not concerned with these categories per se, but rather with quickly locating the relevant words or expressions and their most important senses. Words must not, therefore, be included on the basis of their classical history, grammatical status, or raw frequency, nor must they be omitted merely because they are derivations or multiword expressions.

By thus recalibrating our ideas of "words" in Arabic and the criteria necessary for judging "importance," we can create reference works in which users and learners can easily find the words and expressions that truly reflect modern usage.