



Towards a Corpus-based Collocations Dictionary of Modern Standard Arabic

Sattar Izwaini

American University of Sharjah

ABSTRACT: This paper discusses the need to compile a collocational dictionary of modern standard Arabic. It reports a project under way of producing a corpus-based dictionary of Arabic collocations. It outlines the corpus composition, the electronic tools used, and the guidelines followed in the carrying out the project. The paper describes the evolution and progress of the project. It will also discuss other linguistic features of words and their collocations such as number, gender, and syntactic configurations.¹

Keywords: Collocations, Arabic, lexicography, corpus linguistics.

1. Introduction

The present paper provides an account of compiling a specialized dictionary, a modern dictionary of Arabic collocations by using modern technology to count for the words and their collocational patterns. A corpus of electronic texts is used to identify words combinations. A corpus is a large set of texts in electronic format (millions of words) that can be processed using certain software.

In order to produce natural idiomatic cohesive texts, native speakers of the language or otherwise, need to have an up-to-date reference that provides them with the established word combinations along with information about their connotations whether positive or negative.

With words that need to combine with each other according to the idiomaticity of the language, a language user, especially students, translators and novice writers, may not be sure which verb, adjective or adverb should be used with a certain word. They need to consult a reference book which is usually a dictionary. For English, for example, there are two well-known collocation dictionaries that provide such information: the BBI Dictionary of English Word Combinations (Benson et al. 1997) and Oxford Collocations Dictionary (2012). There is also Macmillan Collocations Dictionary (Rundell, 2006). Arabic collocation dictionaries will be outlined below.

After this introductory section, the next section will provide a brief outline of Arabic lexicography and Arabic collocation dictionaries, classical and modern. Section three will discuss collocations in Arabic and its types. Section four will provide a discussion of the corpus and tools used.

Section five will present the dictionary design. Examples will be given to clarify how these combinations are to be identified and listed, accompanied by their transliterations. Literal translations of examples are provided in brackets. Section six will review the methodology used. The database created will be addressed in section seven. The final section is the conclusion which will provide an outline of the paper as well as suggestions for future work.

¹ This paper represents the opinions of the author and does not mean to represent the position or opinions of the American University of Sharjah.

2. Arabic lexicography

Arabic lexicography goes back to the eighth century with dictionaries were compiled to serve different purposes. Starting with monolingual general dictionaries, Arabic lexicography developed and progressed to produce general and specialized dictionaries. There have been also dictionaries that are concerned with a certain vocabulary or terminology of a specific field. In modern times, bilingual and trilingual dictionaries whether general or specialized have been published.

2.1 Classical collocational dictionaries

Arabic classical dictionaries listed the lexical component of the language covering words, collocations and idiomatic expressions in general. The most well-known collocational dictionaries are *Al-mukhassas* المخصص (the specialized) by Ibn Seedah (11th century/1970) and *Fiqh Al-lugha* فقه اللغة (Arabic Philology) by Al-thaʿaaliby (11th century/1996). While the majority of entries of these dictionaries are collocations, they also include other idiomatic expressions such as idioms, couplets (*itbaa* اتباع in Arabic) and multiword expressions. These two classical collocation lexicons are about one thousand years old. They include obsolete usages. This also applies to general classical Arabic dictionaries that have collocations within their entries. There is, therefore, a need to compile a modern dictionary that excludes such usages.

2.2 Modern collocational dictionaries

Al-yaazijy (1904/1985) wrote a dictionary of phraseology that covers a wealth of collocations as well. In the present time, two dictionaries of Arabic collocations have been published. The first is Hafiz Collocations Dictionary (2004) (معجم الحافظ للمتصاحبات اللفظية) which is an Arabic-English, non-corpus-based containing 7756 collocations. The second is Dar El-ilm Dictionary of Collocations (Ghzalah, 2007) (قاموس دار العلم للمتلازمات اللفظية). It is a bilingual English-Arabic, with mostly idiomatic expressions including collocations and non-collocations. These two dictionaries are either based on intuition and/or other non-corpus-based methods. They list collocations along with idioms and fixed expressions as no difference is made between these categories. Dyaab (1996) published a monolingual Arabic dictionary of sounds which is a restricted area of collocations.

The dictionary that is discussed in this paper has been planned to be corpus-based where evidence of actual language use is recorded. Its text material is available in electronic format to be processed, and collocations to be traced and documented. The dictionary also is significant in recording new collocations whether produced by the need for new word combinations or as an outcome of translations from foreign languages (more discussion of new collocation will follow).

3. Collocation

Collocation is an idiomatic manifestation of language use next to couplets, idioms and multi word expressions. It is the frequent and habitual co-occurrence of words where they constitute an idiomatic combination. In English, for example, it is *strong tea* not **powerful tea* nor **dense tea*. Words that describe good or bad conditions of things are different although they refer to the same concept: *immaculate performance*, *flawless argument*, *impeccable behaviour*, and *spotless reputation*. There are also *rancid butter* and *addled egg* with different adjectives that mean the same, but they are not interchangeably used with those head words (nouns), i.e. not **addled butter* nor **rancid egg*. In addition, groups are differently designated: *school of fish*, *pack of wolves*, and *flock of birds*. Specific verbs are used with specific nouns: *to wear perfume* and *to pay a visit*.

Languages have different collocational patterns, and language users may not know which specific verb or adjective is to be used with a certain noun. Translators can have difficulty in dealing with foreign collocations while translating texts. A simple example is the Arabic counterpart of *strong tea* is *shaayun thaqeel* شاي ثقيل (literally heavy tea). However, 'heavy' is used in English to describe *rain* (among many other words) whereas in Arabic *thaqeel* ثقيل (heavy) is not, but rather it is *matarun ghazeer* مطر غزير (pouring rain).

Depending on the range of words collocating with a certain word, the collocations result of the combination can be said to be bound like *مسقط رأس masqatu ra's* (place of birth) where *مسقط masqat* has one collocate only. It is restricted collocation where the number of collocates is limited, or open collocation where the range of is so extensive (Cowie, 1981: 228).

In Arabic, couplets are idiomatic expression with no collocation combinability of their elements as the case is with collocations. Their main purpose is emphasis and stylistic effect. Couplets have a special idiomatic nature similar to bound collocations. They are so rigid in that they are close to idioms, but they are transparent like collocations.

A couplet can incorporate two words attached by a coordinator. There can be no coordination involved. Its elements can be synonymous, semi-synonymous, or opposite words, for example *الصبر والسلوان al-sabru wa al-sulwaan* (patience and solace) and *خرايب يباب kharaab yabaab* (devastated and desolate). They can be derived of the same root as in *القال والقال al-qeel wa al-qaal* (hearsay and gossip). Some couplets have two versions with their element can be either in a verb, noun or adjectival form, e.g.:

Adjective form	Verb form
كأفافية وافية	كفأف وواف
<i>kaafiyatun waafiyyah</i>	<i>kaffaa waffaa</i>
(achievement and accomplishment)	(achieved and accomplished)

Idioms are expressions whose meaning is not the total meanings of its constituent elements. They are closely related to the culture of the language. Idioms are not transparent and need to be looked at and processed as one semantic unit, for example *دونة خراط القتاد doonahu khartu al-qataad* (literally: it would be easire to collect leaves or sticks from a thorny bush) (idiomatically: impossibly attainable).

4. Collocation types

Collocations are to be identified and categorized in accordance with the categories proposed by Izwaini (2015). He divided collocations according to words classes of nouns, verbs, adjectives and adverbs. There are sub-categories that involve other word classes. Below is a summary of the main patterns of Arabic collocations. For a detailed account, see Izwaini (2015).

1. Noun–Noun (n - n): an additive construct (a noun and a modifier) such as *حفاف الافاف hafeefu al-af'aa* (rustle of a snake), *قطة ذئاب qate'u thi'aab* (pack of wolves), *مقر الحكومة maqarru al-hukoomah* (government headquarter).
2. Noun–Adjective (n - adj): *فرصة سانحة fursatun saanihah* (available opportunity), *زي رسمي zay rasmy* (formal attire).
3. Noun–Verb – (n - v): an intransitive verb and a noun (subject): *باز القمر bazaga al-qamar* (the moon rise).
4. Adjective–Noun (adj - n): *طيب القلب tayibu al-qalb* (kind-hearted). The adjective *آيل 'aaiyl* (definitely going to) has one collocation only which is this particular noun *آيل للسقوط 'aaiyl lil-suqoot* (prone to collapse).
5. Adjective–Adverb (adj + adv) (absolute object): *سريع هائلة saree'un surh'atan haa'ilah* (speedy a tremendous speed). This is what I call a compound collocation (see discussion below)
6. Verb–Noun (v - n): an transitive verb and a noun (object): *أفقدني أفقد afqadany sawaaby* (drive me crazy).
7. Verb–Adverb (v - adv): *سار سريعاً saara musri'an* (to walk quickly).

While collocations in general can be simple with two or more elements (words), they can also be compound with more than one collocation included in one larger combination. A compound collocation is where two

collocations incorporated together. It includes a combination of two words collocating with a third one. For example, *drasahu dirasatan mustafeedhah* (to study it thoroughly), the collocation *دراسة مستفيدة*. The noun here collocates with the adjective. This collocation is embedded in the whole collocation *درسه دراسة مستفيدة* in which the verb collocates with the noun- adjective collocation.

5. Corpus and tools

I will discuss here the make-up of the corpus with all its components as well as its size in terms of number of words. This section will also review the tools used to process the corpus.

The corpus is organized in folders of sub-corpora to be processed. The tracing of word combinations was embarked on by processing the texts using the Wordsmith tool. The identified collocations have been entered into a database that was specially created for this purpose. The entries were organized with one sheet of the database for each letter of the alphabet.

The available corpus is about 3 million words of four classical general-use dictionaries:

1. Al-kaamil (Al-mubarid, 3rd century/1997).
2. Al-sihaah (Al-Jawhary, 4th century/1990).
3. Lisan Al-arab (Ibn Manzoor, 7th–8th century/1999)
4. Al-qaamoos Al-muheet (Al-fayroozabady, 9th century/2005).

Both collocational dictionary of *Al-mukhassas* (the specialized) by Ibn Seedah (11th century/1970) and *Fiqh Al-lugha* (Arabic Philology) by Al-thaaliby (11th century/1996) are included in the corpus. The corpus also includes one pre-modern dictionary of Arabic phraseology, *Naj'at Al-raa'id wa Shar'at Al-wa'rid fy Al-mutradif wa Al-mutawa'rid* (Al-yaazijy, 1904/1985).

The data incorporates an open-source sub-corpus of modern Arabic newspapers from different Arab countries with diverse topics and areas (art, education, finance, health, music, religion, sports, politics, and science & technology). This includes four sub-corpora of modern Arabic:

1. Corpus of Arabic by M. Abbas of 6000 news articles of *Al-khaleej* and *Al-watan* newspapers (politics, economy, health, religion, sports, finance, education, news science & technology, and art & music) (Abbas et al. 2011).
2. Kalimat: a multipurpose corpus (4 million words) by M. Elhaj, which includes Omani newspaper *Alwatan* (20,291 Arabic articles).
3. Aracorp by A. Abdali: 140 million. It includes the news papers: *Al-akhbar*, *Al-watan*, and *Al-tajdeed*.
4. Tashkeela Arabic Corpus (75 million words) (Tashkeela) (see also Zerrouki and Balla, 2017). It is a collection of Arabic vocalized texts, which covers modern and classical Arabic language.

To have as a comprehensive coverage of text types as possible, it has been also planned to use online searchable corpora such as ArabiCorpus (173 million words), and the International Corpus of Arabic (100 million words).

To process the data, corpus applications have to be used. One of these applications is Wordsmith which is a suite of tools that generate word list, search for word combinations among others. Also, there is Sketch Engine which is web-based tool that allows processing a corpus the way intended to harvest the lexical combinations. ArabiCorpus is a web-based corpus and tool that helps in searching for words in their respective contexts whether a noun, a verb, adjective, or adverbs, or even strings of words, to be looked for to identify their collocations.

6. Dictionary design

The planned dictionary is corpus-based where evidence of actual language use is available in electronic

format to be traced and documented. The dictionary is devised to provide an up-to-date source as which word needs to be used with which word, and will specify the word connotation whether positive or negative. The dictionary is designed to offer a specialized language reference needed by students, language instructors, translators and writers as well as general readership.

This project of modern collocation dictionary started first with a manual part that included surveying modern hard-copy general and collocation dictionaries such as *Al-'alfaaz Al-kitaabiyyah* (Al-hamadaany, 10th century/1980). The collocations are recorded in a database that was especially created for this purpose. Collocations found in those dictionaries are listed according to their alphabetical order.

The collocations that have fallen out of use had to be removed. On the other hand, many new collocations have emerged with words assuming new combinations that need to be recorded and included in this new dictionary. For example, the collocation denoting a group of airplanes, *sirb taa'iraat*, *sirb* involves the word *sirb* that is initially used for birds as in *sirb tuyoor* (flock of birds). Another example is the collocation that describes the noise of a tank *hadeer al-dabbabah*, which borrowed the sound of either the camel *hadeer al-ba'eer* or sea waves *hadeer al-bahr*.

With new concepts and terms finding their way into Arab culture and language, new collocations have been created to convey these concepts and terms. Below are some examples

أجرى مفاوضات (v. n. obj.) *ajraa mufaawadhaat* for to conduct negotiations.

خاض الانتخابات (v. n.) *khaadha al-intikhabaat* for running for elections. التصبّأت

امتنع عن (v. prep. n.) *imtana'a an al-tasweet* for to abstain.

نقل مباشراً (adj. n.) *naql mubaashir* for live transmission.

Another kind of new collocations in Arabic is the calque (loan) translations of foreign expressions. Translation has played a major role in providing Arabic with new collocations that reflected collocations of foreign languages, e.g. *tanmiyah mustadaamah* for *sustainable development* and *ghasl al-'amwaal* for *money laundering*. Naturally, such Arabic collocations are not necessarily found in specialized or general dictionaries. It is, therefore, clear why language users need an updated lexicon of collocations that can be consulted while writing, translating or carrying out research. While writing, Arabic language users' knowledge, intuition and/or memory may fail them in deciding on the word combination needed for a specific concept.

The second stage is the computerized processing of the corpus to trace the lexical combinations to be recorded and included in the database. The collocations are categorized according to their types, structures and restrictions. Issues such as semantic features and syntactic configurations of collocations are highlighted.

7. Methodology

This section will discuss the methodology used. First lexical combinations are identified, excluding archaic combinations. Collocations are extracted and documented, organizing entries by nouns first then moving to other word classes such as verbs, adjectives and adverbs, and cross-recording collocations.

Words are checked to find their word combinations starting from the first letter of the alphabet moving on a letter by letter. The patterns of collocations such as noun-noun, adjective-noun, verb-noun, and verb-adverb as well as compound collocations are recorded. The entries are organized and their connotations to be indicated.

Cross reference of collocations of other letters involved in the combination have to be made at the same time to document them in their own database sheets. Otherwise, such collocations could be missed and not recorded. They would not be identified later which can be a flaw in the project. This step has to be taken while working on the other element of the collocation, making the time allocated for each letter longer than planned.

The semantic prosody of collocation, i.e. whether the collocation has a positive or negative connotation

(Louw, 1993), is also to be highlighted, by listing collocations starting with the positive ones, moving to neutral and finally the negative ones. To give an example of this aspect of collocations, the verb يَكُنُّ *yakunn* (to have a feeling) has a positive prosody, because it tends to collocate with nouns as its object like احْتِرَامٌ *ihтираam* (respect), تَقْدِيرٌ *taqdeer* (appreciation), and حُبٌّ *hubb* (love). On the other hand, the verb يَضْمُرُ *yadhmir* (to have a feeling) has a negative prosody as it collocates with words like شَرٌّ *sharr* (viciousness), ضَغِينَةٌ *dhageenah* (malevolence), and حَقْدٌ *hiqd* (malice). These collocates are not used interchangeably. They do not combine with each other's collocates (see also Al-tha^caaliby, 1996: 352 regarding the negative prosody of the verb هَاجَ *haaja* (erupt), Louw, 1993: 157, and Izwaini, 2015: 79).

After having the corpus organized, the electronic processing was embarked on using the software available as well as internet-based corpus and online tool have been used. Some challenges that are faced are the non-availability of some internet-based corpora. Institutions and researchers may opt for not making corpora available online in a later stage after the launching of the corpus. Another challenge is difficulties that can be faced when using online searchable corpus like the International Corpus of Arabic (100 million words). It has not been helpful as it does not yield useful results and is not user-friendly. It takes very long time when searched as well. Therefore, ArabiCorpus (173 million words) has been extensively used to trace collocations.

To electronically process the corpus of texts to identify and document word collocations, a database has been prepared with entries for each of the letters of the Arabic alphabet. The database has evolved with entries of word combinations added and cross-referenced. It is envisaged that the processing of the corpus should be as comprehensive as possible to cover most if not all word combinations of Arabic.

I must also highlight the cross-recording of the components of word combination. For example, a head noun and a modifier noun can both be head and collocate. A noun collocating with another noun can be the collocate of the latter. As the initial plan was to process each letter of the alphabet and when finished, I would move to the next letter. However, when one collocation was identified and entered under the letter of the head word, the same collocation had also to be documented under the letter of the other word. For example, the collocation مَطَرٌ غَزِيرٌ *matarun ghazeer* (heavy rain) has to be entered under the first letter مِيمٌ *meem* as well as under the letter غَايٌ *ghayn* for the word غَزِيرٌ *ghazeer*. Moreover, all collocations with the word *ghazeer* غَزِيرٌ such as شَعْرٌ غَزِيرٌ *sha^crun ghazeer* (literally abundant hair) should be entered under the letter شَيْنٌ *sheen* as well.

A similar case is when some words function as a headword in one collocation as well as a collocate in another. For example, the word سَوْءٌ *soo'u al-sulook* (misconduct literally bad behaviour) as well as the modifier collocate in ظَنُّ السَّوْءِ *zanu al-soo'* (dark suspicion). In this case, the dictionary has to accommodate both collocations with separate entries for each.

The entries of verbs need to be in the past tense to avoid having all verbs being under one letter of *yaa'* which the present form prefix at the beginning of the word, e.g.

Verb	Alphabet entry	Present form	Alphabet entry
كَتَبَ <i>kataba</i> (to write)	ك (k)	يَكْتُبُ <i>yaktub</i>	y
جَاءَ <i>jaa'a</i> (to come)	ج (j)	يَأْتِي <i>iyay</i>	y
فَرِحَ <i>fariha</i> (to be glad)	ف (f)	يَفْرَحُ <i>yafrah</i>	y

Derived forms of words with prefixes (أَحْرُفُ الزَّيَادَةِ *ahruf al-zyaadah*) need to be listed in the dictionary as separate entries since users may not know how to derive them from the stem like the verbs أَلْهِمَ *alham* (to inspire) and اسْتَلهِمَ *istalhama* (to get inspiration).

The dictionary entries should also consider word senses. Words can have more than one sense. For example, one sense is literal and tangible as in صُورَةٌ مَزَيَّافَةٌ *suratun muzayafah* (forged picture), and the other is metaphorical as in صُورَةٌ زَائِفَةٌ *suratun zaa'ifah* (fake image). A similar case is the collocations of بَرَاءَةٌ *baraa'ah* whose collocation بَرَاءَةُ الْأَطْفَالِ *baraa'atu al-'atfaal* (children's innocence) is of general profile while its

collocations براءة اختراع *baraa'atu ikhtiraa'* (patent) and براءة المتهم *baraa'aut al-mutaham* (literally the accused's innocence) have a specialized profile that needs to be listed under different sub-entries indicating their specialization.

Another example is the term تلقيح *talqeeh* which can have two senses, one for the process of flower pollination تلقيح الأزهار *talqeehu al-azhaar*, and the other denotes vaccination as in تلقيح من الأمراض *talqeehun mina al-amraadh* (vaccination). In this case the word تلقيح *talqeeh* should have two sub-entries.

A case in point is whether specialized terms need to have entries in the dictionary. This may lead to have a very large work beyond the boundaries of general language vocabulary. It will move into the realm of special areas rather than be dedicated to general-language use. Then the decision should be excluding terminology. It is a specialized vocabulary of fixed meaning and formulation that make it of a specific use and unique character. It does not have the variability of general vocabulary. However, the borderline is fuzzy as many specialized vocabularies find its way into the general language and have some traits of collocations. Terms are widely used by language users in everyday life contexts such as تلقيح من القولنج (n. prep. n.) *al-talqeeh mina al-koleraa* for Cholera vaccination.

Prepositions can be indispensable for many collocations (see categories in section 3 above). They are inherent grammatical and syntactic elements that part and parcel of many collocational expressions, for example بريء من التهمة *baree'un mina al-tuhmah* (literally innocent of the charge). They have to be stated in the entries and sub-entries to help the dictionary user know which preposition to be used in one particular collocation. However, prepositions should not have separate entries as they are not part of the language lexis.

In many cases, a word can have collocations with both the singular as well as plural forms. Plural collocates need not to be listed as users should be able to decide whether it is the singular form, or the collocation is of the plural form rather than the singular following the language syntactic rules as in براءة الأطفال *baraa'atu al-'atfaal* (children's innocence) which is of a generic reference rather than the singular form براءة الطفل * *baraatu al-tifl* (child's innocence). The latter refers to a particular baby. An example collocation of a plural node is إجراء مراسم *ijraa'u maraasim* (holding ceremonies).

The entries are of nouns in the first place. In case it is a noun-noun collocation, both constituents should be recorded under their respective alphabet letters. The entry starts with nouns, followed by adjectives and then by verbs. However, a verb collocate can be related and derived from the noun collocate anyway. Verbs of the same root are most likely to be used with the same collocate of their corresponding nouns:

<u>Noun collocation</u>	<u>Verb collocation</u>
التعرض للإشعاع <i>al-ta'arudhu lil-ish'aa'</i> (exposure to radiation)	تعرض للإشعاع <i>ta'aradha lil-ish'aa'</i> (exposed to radiation)
الشفاء من المرض <i>al-shifaa'u mina al-maraadh</i> (recovery from illness)	شفي من المرض <i>shafiya mina al-maradh</i> (recovered from the disease)
تأني المخدرات <i>ta'aaty al-mukhadiraat</i> (taking drugs)	تأني المخدرات <i>ta'aataa al-mukhadiraat</i> (took drugs)

In some cases, nouns can be adjectives as well and vice versa, cf. عاصفة *aasifah* (storm) and رائج عاصفة *reehun aasifah* (stormy wind). The latter in this case is a collocate of رائج *reeh* (wind) whereas the former is a noun and will have its own entry.

Assuming the user knows the basics of the language such as the definite and indefinite forms of nouns like those words so called مقصور *maqsoor* (have a final long vowel), e.g. with a final الف *alif*, ياء *yaa*, or واو *waaw* to be removed when indefinite and kept when having the word is prefixed with definite article *al*. For example ناد *naadin* (club) vs. النادي *al-naady* (the club). Similarly, and based on the same rationale, feminine

forms of nouns or adjectives will not be provided only for exclusively feminine gender like حامل *haamil* (pregnant) and مريض *murdhi* (breast feeding).

As well, the user is expected to have the ability to adapt a noun-adjective collocation to be an adjective-noun one, making a compound adjective, so called إضافة غيحية *idhaafatun ghayru haqeeqiyyah* (non-actual construct), where this collocation acts as a compound adjective modifying a noun:

<u>Noun-adjective collocation</u>	<u>Adjective-noun collocation</u>
بديهة <i>badeehatun saree^{ah}</i> (literally quick wit)	آع البديهة <i>saree^u al-badeehah</i> (literally quickly-witted)
رأس مرفوعة <i>ra'sun marfoo^{ah}</i> (head held high)	مرفوع الرأس <i>marfoo^u al-ra's</i> (literally of a held-high head)

The dictionary should also record new collocations whether produced by the need for new word combinations as in طائرات *sirbu taa'iraat*, (squadron of airplanes) or as an outcome of translations from foreign languages like علاقات ثنائية *alaaqaatun thunaa'iyyah* (bilateral relations).

Although the dictionary is corpus-based, some intuition needs to supplement the findings. The lexicographer's knowledge of the language system and its lexical component should enable him/her to decide whether the combination follows the language rules and stylistic norms as the case of many awkward and wrong usages of prepositions involved in collocations or opting for the adjective-noun collocations rather than the canonical noun collocations (see also Izwaini, 2014).

Due to not having a good command of the language, many users do not opt for the noun-noun collocation rather than its adjectival counterpart which does not comply with the sound Arabic syntax, for example, اشارة تحذيرية *ishaaratun tahtheeriyyah* (warning adj. sign) (n – adj.) instead of the noun – noun اشارة تحذير *ishaaratu tahtheer* (warning sign) where the latter should be adopted. This also applies to other ungrammatical structures or awkward styles that do not follow the language norms and canonical structure.

Also, as a corpus would provide how the language is used, some caution should be exercised. Prescriptive approach should be used to help avoid awkward non-standard (*fusha*) style that is produced due to interference from foreign languages directly or via translation, and due the fact that standard language is not always mastered by language users.

8. The database

The database was designed to have a chapter for each alphabetic letter, recording all words starting with that letter according to word classes: nouns, adjectives, verbs, adverbs and their collocates, including prepositions involved as outlined above.

Collocates of different senses are listed in groups clearly identified and clustered, starting with collocations of favourable/positive senses, then those with neutral senses and ending with those of unfavourable and/or negative ones. Collocates are to be listed in groups of one semantic field or levels of favourable unfavorable (see Oxford Collocations Dictionary, 2012, and BBI collocation dictionary, Benson et al. 1997).

8.1 Entry sample

Here, a sample is provided to demonstrate the collocational range of the noun فرصة *fursah* (opportunity) according to the dictionary design and arrangement.

- Noun-noun

مئح، اتاحة، اغتنام؛ اقتناص، تحي؛ فن؛ انتهاز، استغلال، اضاعة، تفانت (الفرصة) ع؛ ضاع،
توفر؛

- Noun-adjective

ثمينة، عظيمة، ذهبية، سائحة، مواتية، متاحة، طيبة تاريخية؛ أحسن، أفضل، أقر، ممكنة، مناسبة؛ ضرة.

- Verb-Noun

A- Subject

أسح، تحي، فن، نوات، تضع (ع)، تفت؛

B- Object

أمنح، يتيح، أغتنم؛ ينهز، أقتنص، يتحي، فن، استغل؛ أضيع، أفوت (الفرصة) ع.

9. Conclusion

This paper has discussed the project of compiling a corpus-based dictionary of collocations of modern standard Arabic. It underlines the rational to compile such dictionary. This includes the removal of out-of-date collocations, and incorporate new word combinations that entered the Arabic language in modern times. The paper has also discussed the corpora and tools used to identify lexical combinations, collocational types to be traced, and the guidelines that have been followed in detecting and recording collocations. It also outlined the challenges that have been faced while working on this project.

For future work, bilingual collocational dictionaries can be planned to incorporate Arabic-English collocations to be used by translators in particular.

References

- [1] Al-fayroozabady, M. (9th century/2005). *Al-qaamoos Al-muheet* (the unbridged dictionary). Beirut: Mu'ssasat Al-risaalah.
- [2] Al-hamathaany, A. (10th century/1980). *Al-'alfaaz Al-kitaabiyyah* (الألفاظ الكتابية) No place: Al-daar Al-'arabiyyah lil-kitaab.
- [3] Al-jawhary, I. (4th century/1990). *Al-sihaah* (الصاح). Beirut: Dar El-Ilm Lil Malayin.
- [4] Al-mubarid, M. (3rd century/1997). *Al-kaamil* (الكمال) (the perfect). Cairo: Daar Al-fikr Al-arabiyy.
- [5] Al-tha'aaliby, A. (11th century/1996). *Fiqh Al-lugha wa Sirr Al-arabiyyah* (فقه اللغة و السر المكتوبة) (Arabic philology and the secret of the Arabic language). Beirut: daar Al-kitaab Al-arabiyy.
- [6] Al-yaaziy, I. (1904/1985). *Naj'at Al-raa'id wa Shar'at Al-waarid fy Al-mutraddif wa Al-mutawaarid* (نجعة الرائد و شارة الوارد في المصنفات والمترادفات) (the resource of collocations and word combinations). Beirut: Librarie du Liban.
- [7] ArabiCorpus. <http://arabicorpus.byu.edu/>
- [8] Benson, M., Benson, E. and Ilson, R. (1997). *The BBI Combinatory Dictionary of English Word Combinations*. Amsterdam: John Benjamins.
- [9] Cowie, A. P. (1981). "The Treatment of Collocations and Idioms in Learner's Dictionaries." *Applied Linguistics*, 2(3): 223-235.
- [10] Dyaab, K. (1996). *Al-mu'jam Al-mufasssal fy Al-'aswaat* (المعجم المفصل في الأصوات) (the detailed dictionary of sounds). Tripoli: Gross Press.
- [11] Ghazala, H. (2007). *Dar El-Ilm's Dictionary of Collocations* (قاموس دار العلم للمتلازمات اللفظية) Beirut: Dar El-Ilm Lil Malayin.
- [12] Hafiz, A. (2004). *Hafiz Collocations Dictionary*. (معجم الحافظ للمتصاحبات اللفظية) *Mu'jam Al-hafiz lil-mutasaahibaat Al-lafziyyah*. Beirut: Librarie du Liban.
- [13] Ibn Farish, A. (4th century/1947) *Al-itbaa' wa Al-muzaawajah* (الاتباع والمزاوجة). Baghdad and Cairo:

- Maktabt Al-muthanaa and Maktabath Al-khanjy.
- [14] Ibn Manzoor, M. (7th–8th century/1999): *Lisaan Al-ʿArab* لسان العرب (the tongue of Arabs). Beirut: Daar ihyaa' Al-turaath Al-ʿaraby and Mu'sassat al-tārīḥ al-ʿaraby.
 - [15] Ibn Seedah, A. (11th century/1970). *Al-mukhassas* المخصص (the specialized dictionary). Beirut: Al-maktab Al-tijaary.
 - [16] International corpus of Arabic. <http://www.bibalex.org/ica/ar/>
 - [17] Izwaini, S. (2015). Patterns of Lexical Collocations in Arabic. *Zeitschrift für Arabische Linguistik*, 61: 71-99.
 - [18] Izwaini, S. (2014). Translation and Lingusitic Recycling. *Bable*. 60: 4, 478-513.
 - [19] Kalimat. <https://sourceforge.net/projects/kalimat/>
 - [20] Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker M., G. Francis and E. Tognini-Bonelli (eds.), *Text and Technology: in Honour of John Sinclair*. Amsterdam: John Benjamins. 157-176.
 - [21] Oxford Collocations Dictionary. (2012). Oxford: Oxford University Press.
 - [22] Rundell, M. (2006). *Macmillan Collocations Dictionary*. Macmillan ELT.
 - [23] Sketch Engine. <https://www.sketchengine.eu/>
 - [24] Tashkeela: Arabic Vocalised Texts Corpus. <https://sourceforge.net/projects/tashkeela/>
 - [25] Wordsmith tools. <https://www.lexically.net/wordsmith/>
 - [26] Zerrouki, T. and Balla, A. (2017), Tashkeela: Novel corpus of Arabic vocalized texts. <http://dx.doi.org/10.1016/j.dib.2017.01.011>