UNIVERSITY OF LEEDS

This is a repository copy of *Constructing a corpus-informed list of Arabic formulaic sequences (ArFSs) for language pedagogy and technology*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/144498/

Version: Accepted Version

Article:
Alghamdi, A and Atwell, E orcid.org/0000-0001-9395-3764 (2019) Constructing a corpus-informed list of Arabic formulaic sequences (ArFSs) for language pedagogy and technology. International Journal of Corpus Linguistics, 24 (2). pp. 202-228. ISSN 1384-6655

https://doi.org/10.1075/ijcl.16088.alg

(c) 2019 John Benjamins Publishing Company. This is an author produced version of a paper published in International Journal of Corpus Linguistics. Please contact the publisher (John Benjamins) for permission to re-use or reprint this material in any form. Uploaded in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Constructing a corpus-informed list of Arabic formulaic sequences (ArFSs) for language pedagogy and technology

Ayman Alghamdi and Eric Atwell

Umm Al-Qura University | University of Leeds

This study aims to construct a corpus-informed list of Arabic Formulaic Sequences (ArFSs) for use in language pedagogy (LP) and Natural Language Processing (NLP) applications. A hybrid mixed methods model was adopted for extracting ArFSs from a corpus, that combined automatic and manual extracting methods, based on well-established quantitative and qualitative criteria that are relevant from the perspective of LP and NLP. The pedagogical implications of this list are examined to facilitate the inclusion of ArFSs in the process of learning and teaching Arabic, particularly for non-native speakers. The computational implications of the ArFSs list are related to the key role of the ArFSs as a novel language resource in the improvement of various Arabic NLP tasks.

Keywords: lexical resources, Arabic formulaic sequences, multi-word expressions, language pedagogy, mixed methods

1. Introduction

The phenomenon of multi-word expressions (MWEs) in human language has attracted the attention of researchers in various language-related disciplines e.g. linguistics, psychology, language pedagogy (LP) and Natural Language Processing (NLP). Hence, this phenomenon has been researched from a number of different scientific angles. A considerable amount of research has evidenced the major role of MWEs in the process of analysing, learning and understanding languages. From a linguistic perspective, many studies have emphasised the crucial importance of including formulaic language and MWEs in second language learning and teaching. Several researchers have highlighted the fact that the mental lexicon

is not merely represented by single orthographic words, but rather it incorporates longer formulaic sequences (FSs) (e.g. Pawley & Syder, 1983; Kjellmer, 1990; Wray, 2002). Other researchers have attempted to develop MWEs lists, which can be used as a pedagogical tool in language teaching and learning e.g. material design, curriculum developments and language testing. On the other hand, from a computational perspective, MWEs play a vital role in NLP and many researchers have attempted to construct various types of MWEs repositories in order to integrate them in the development of various NLP software systems (e.g. MWEs identification and extraction, language Part-of-Speech tagging and parsing, information retrieval and named entity recognition).

The vast majority of research in this area has been conducted with the English language because of the interest in and demand for English language teaching, and the rich availability of free access English language resources. Recently, Arabic has received increasing attention from researchers from different, albeit related, disciplines. However, in comparison to English, Arabic MWEs research is still at an early stage. The key role of formulaic language and MWEs resources in LP and NLP and the lack of free access to Arabic MWEs lexical resources are drivers for research on constructing an Arabic corpus-informed MWEs list for LP.

The main objectives of our study are twofold:

i. A guide for Arabic language learners and educators to include ArFSs in their learning and teaching, particularly for non-native speaker learners.
ii. A comprehensive computational corpus-informed ArFSs lexical resource, which can be incorporated into various Arabic NLP applications.

In this paper, we report on empirical research to develop and apply a hybrid model for extracting ArFSs from a corpus. The paper is organized as follows. Section 2 discusses definitions of FSs, and related work from the linguistic and computational perspectives. Section 3 presents the empirical methodology. Sections 4 and 5 present the empirical procedure and the results of adopting a hybrid model for extracting ArFSs from a corpus. Finally, we draw conclusions in Section 6.

2. Formulaic Sequences in language pedagogy and technology

When attempting to define the FS, the heterogeneous nature of this phenomenon in human languages at different linguistic levels can be clearly noticed, e.g. morphology, syntax and semantics. Hence, it is hard to find a consensus in the literature on what we can call FSs. This is mainly due to the complexity involved in the linguistic properties of FSs, like the well-known tale about blind men feeling different parts of an elephant and each giving a different description, every researcher attempts to demonstrate his or her own understanding of this complicated phenomenon. For instance, in Computational Linguistics and NLP the term 'multi-word expression' (MWE) is used to refer to various linguistic items including, but not limited to, idioms, noun compounds, phrasal verbs and light verbs (Sag et al., 2002; Gralinski et al., 2010). Hence, a precise, complete and comprehensive definition of FSs is beyond the reach of our study, particularly in morphologically rich languages as is the case in Arabic. Because of this, a practical definition will be suggested for this study, which defines the types of FSs targeted in the current research. This definition is based on our research objectives that mainly focus on Arabic expressions that are most useful for pedagogical uses, particularly phrases that pose difficulty from the perspectives of second language learner comprehension and NLP tasks.

In the literature, many definitions of FSs have been suggested (e.g. Baldwin et al., 2003; Baldwin & Kim, 2010; Ramisch, 2012; Schneider et al., 2014; Wood, 2015). Researchers have specified criteria for recognising or defining FSs in texts and corpora (Leech et al., 2001; Wray & Namba, 2003; Wray, 2009; Schmitt & Martinez, 2012; Wood, 2015). For instance, Wray & Namba (2003) propose a set of eleven criteria that help the researchers to use their intuitive judgment in the manual identification of FSs. These criteria, along with others suggested by previous research (e.g. Coulmas, 1979; Peters, 1983; Wood, 2010a) were considered when developing a set of criteria for this study. The working definition adopted in the current study is based on an integration between two of the most cited definitions of FSs proposed by Sag et al. (2002: 4-5) and Wood (2015: 3). These definitions state the core criteria of FSs which have a consensus in FSs research, and thus here we define ArFSs as: standard Arabic multi-word phrases which have a single meaning or function and present linguistic as well as statistical idiomaticity. This concept of ArFSs covers all types of lexical units that we intend to include in our research because it involves any semantically regular formulas that are not restricted to any syntactic construction or semantic domain. By standard Arabic in our

definition, we exclude other Arabic dialects and focus only on the standard dialect which is the formal type of modern Arabic represented in most forms of communication in the Arabic world today.

Another dimension, regarding the concept of FSs, is the terminology issue. Wray (2002) states that, in the literature, more than 50 terms have been used to refer to this phenomenon; however, Schmitt (2010) suggests the use of the term 'Formulaic Sequence' as an umbrella to refer to various types of FSs in general. Hence, in our study, the term 'Arabic Formulaic Sequences' (ArFSs) will be used because this research covers different sorts of Arabic expressions, and other terms such as MWEs, constructions and collocations might be used interchangeably.

The importance of this research is due to a set of factors related to the vital role of integrating formulaic language in NLP and LP. Ignoring MWEs in any language-related tasks will have a negative impact on their final output quality. This is because MWEs constitute a large part of everyday language; for instance, in English, MWEs constitute 41% of the entries in WordNet 1.7 (Fellbaum, 1998). Li et al. (2003) also state that phrasal verbs constitute approximately one third of the English verb vocabulary. This large portion of MWEs emphasises their key role in the development of language-related applications. Formulaic language research provides evidence that the most frequently used words in languages are only the tip of expressional icebergs (e.g. Martinez & Murphy, 2011). Figure 1 shows the underlying complexity of phrases related to the Arabic word *'ayn* عين ("Eye").[1]

INSERT FIGURE 1 HERE

Figure 1. Tip of the iceberg shows the complexity of phrases related to the Arabic word *'ayn*

Regarding the key role of MWEs in NLP applications, the inclusion of MWE resources can fundamentally improve the quality of many NLP applications, such as computer-aided lexicography, morphological and syntactic analysis, information retrieval, machine translation and foreign language e-learning systems like the Duolingo (http://duolingo.com) and Flax (http://flax.nzdl.org) projects. Integrating FS knowledge in these applications is known to be very beneficial in the reduction of language ambiguity and increasing the accuracy level of NLP system outputs (Ramisch, 2015).

In LP, MWEs play an essential role because they constitute a large proportion of language. Wray (2013), in her timeline for research on formulaic language, finds that research in this area dates back to Firth's famous quote "you shall know a word by the company it keeps" (Firth, 1957: 20). The

early realization of this phenomenon paved the way for many following researchers to conduct empirical and theoretical studies that aimed to have an in-depth comprehension of formulaic language phenomena from different perspectives. In the last two decades, corpus linguistics research findings have demonstrated the essential role of formulaic language (Wray, 2002; Schmitt, 2010). Several examples in the literature highlight the major role of including formulaic language in LP. This is because formulaic language is very common in language; researchers in this area give different estimations of their proportion in language, which ranges from around 30% (Biber et al., 1999) to more than 50% (Erman & Warren, 2000) in spoken and written discourse. Hence, it is difficult to ignore this large percentage of language in any language-related application. Formulaic language also plays a critical role in conveying various kinds of functions and meaning in language communication (e.g. Biber et al., 2004; Dorgeloh & Wanner, 2009; Hyland, 2008; Wulff et al., 2009). For instance, in English and Arabic, many FSs are used as discourse organization signposts (e.g. *man jaha 'aḳrā*, "on the other hand").

Another factor related to the language processing advantages of FSs is the easy acquisition of formulaic language items by native speakers in comparison to non-formulaic items. By contrast, formulaic language acquisition is found to be one of the most challenging and difficult tasks for non-native speakers (Underwood et al., 2004; Siyanova-Chanturia et al., 2011). Other research has emphasized the key role of formulaic language acquisition in the overall improvement of second language learners' proficiency and fluency in the targeted language (Boers et al., 2006). For instance, after analysing the written answers of English as a foreign language (EFL) learners' in a proficiency test, Ohlrogge (2009), finds that students with higher grades use formulaic language more than students with lesser grades.

The final point here is related to the particular importance of Arabic formulaic language research. Many Arabic linguists call for the imperative need for developing different kinds of FSs language resources to utilize them in LP and NLP applications. For example, Omar (2007) and Hawwari et al. (2014) point out the lack of comprehensive Arabic formulaic language resources, particularly in terms of resources that can be integrated easily into LP and NLP applications. Omar (2007) states that most Arabic teaching and learning materials are still based on listings of orthographic single words because of the absence of well-developed FSs resources. Although the importance of English MWEs has been acknowledged by many researchers in the fields of LP and NLP, as evident

by the large number of research papers and dedicated conferences and workshops, the theory of Arabic MWEs is still underdeveloped. There is a critical need for studying Arabic MWEs, both from the theoretical and practical perspectives.

2.1 Corpus-informed pedagogical formulaic sequences

The LOB Lancaster-Oslo-Bergen Corpus was the first corpus aimed at British English language teaching and research (Leech et al., 1983). An early attempt for an automatic extraction of MWEs in English was made by Atwell (1982) in the development of the LOB corpus tagging project. A multi-word or ditto tag list was created for "sequence of two or more orthographically separate 'words' functioning as a signal lexical item" such as *no one* (Leech et al., 1983: 23). This method would be of great importance and usefulness in the automatic identification of immutable phrases from MWEs tagged corpus. However, when it comes to extracting phrases with syntactic and transparent variety, it is difficult to depend completely on automatic extraction methods.

In English, the work of Leech et al. (2001) written during the development of the 100 million-word British National Corpus (BNC) is considered to be the first published attempt to construct a comprehensive corpus-informed phrase list (Martinez, 2011). The generation of this list is based on the result of automatic identification of the most frequently appearing phrases in the POS-tagged written and spoken corpus. The criteria for MWE selection adopted in their research were based on the fixedness of the phrase. Leech et al. (2001: 8) state that "multiword units are items which are treated as a single word token, even though they are spelt as a sequence of orthographic words". For instance, the phrase *so that* was analysed as a single word because it "functions in the same way as a one word conjunction" (Leech et al., 2001: 14). This specification of MWEs forced the researchers to exclude many other important types of FSs such as inflected or separated sequences like phrasal verbs (e.g. *write down* in *write it down*) in that such phrases could not be matched by their POS tagger. This methodology demonstrates some of the limitations of the fully automatic identification of MWEs. It can also be noted that their MWE list is mostly based on frequency data. Statistical data can be very beneficial and accurate in the identification of frequent single orthographic words in contrast to the processing of MWEs that might not provide sufficient information.

Durrant (2009) constructed a pedagogical listing of academic collocations, the main aim being to extend Coxhead's (2000) Academic Word List (AWL) into an academic collocations list to enable integration of formulaic language in LP. The generic list was based on a new English academic corpus developed by the researcher, which includes 25 million academic written words. The development of the list was based on the integration of two approaches: the first is the key collocation approach in which Durrant (2009) aims to find the most relevant word pairs which co-occur with moderate to high frequency within a four-word span across common academic disciplines in the corpus. The second approach is the collocation of academic keywords which aims to find the words that collocate most to the highest-ranked academic key words. Durrant's (2009) approach relies solely on frequency data, and this might lead to ignoring valuable collocation items that do not meet the statistical criteria.

Another study by Martinez & Schmitt (2012) sought to construct a corpus-based list of FSs in general English like the vocabulary general service list (GSL) (West, 1953). The targeted list aims to be used as a pedagogical tool that can be adopted in various pedagogical applications (e.g. language learning, teaching and testing). The list item selection was based on three main criteria related to high frequency, meaningfulness and the non-compositionality of the phrase. A hybrid approach was adopted in the identification of list items, so at the first stage a frequency list of n-gram candidates was generated using WordSmith Tools (Version 6) (Scott, 2016) and the targeted items of a given n-gram were selected manually, based on pre-determined selection criteria. The research yields a list of the 505 most frequent multi-word expressions in English which is called the PHRASE List.

2.2 Arabic computational MWEs research

Arabic computational linguists can enhance Arabic NLP software by accommodating MWEs within language processing For example, Attia (2006) developed an Arabic MWEs list to use in a MWE transducer in the Arabic morphology parser in order to enhance its analysis. Attia used semi-automatic methods of identifying the MWEs to build a list of Arabic MWEs. Based on the classifications of MWEs presented by Sag et al. (2002), Attia classified the Arabic MWEs into four categories related to semantic compositionality and syntactic flexibility. No further details about the corpus used for developing the MWEs list and the selection criteria of the list items is provided by Attia (2006).

Another reason for interest in MWEs in NLP research can be found in text classification. Hawwari et al. (2012) aim to learn how to statistically categorise new MWEs in large corpora. They constructed a general-purpose list of Arabic MWEs, with the list items being compiled manually from different Arabic MWE dictionaries (Abou-Saad, 1987; Seeny et al., 1996; Dawod, 2003; Fayed, 2007). The final list amounts to 4,209 MWEs. The list was then automatically tagged with the parts-of-speech tagger *MADA* (Habash & Rambow, 2005). The MWEs were manually classified by their syntactic constructions, e.g. Verb-Verb, Verb-Noun, Verb-Particle, Noun-Noun and Adjective-Noun. They developed a pattern-matching algorithm for classifying text in Arabic corpora. The pattern-matching algorithm was run on the Arabic Gigaword 4.0 corpus (AGW) to tag the Arabic text automatically with MWE annotations. The manual evaluation of a sample of automatic MWE annotations reveals an encouraging result with a high degree of accuracy.

Another study by Hawwari et al. (2014) presents a framework for classifying and annotating Egyptian MWEs. The research sought to build an intensive lexical resource for dialectal Egyptian Arabic NLP, enriched with comprehensive linguistic annotations which include phonological, orthographic, semantic, morphological, syntactic and pragmatic information. The list is composed of 7,331 MWEs compiled from corpora and MWE dictionaries.

Although computational linguistics research into MWEs has been applied to Arabic language texts, as is the case in our study, our study has different objectives. The above examples illustrate Arabic MWE lists developed to improve Arabic NLP software; to the best of our knowledge, no Arabic MWEs list has been developed for language pedagogical purposes. Hence, our study seeks to fill in the gaps in our knowledge by developing a pedagogically-relevant and corpus-driven list of MWEs.

3. Methodology: A hybrid model for FSs extraction

Our mixed methods model for extracting FSs aims to combine statistical methods with qualitative methods; hence, quantitative and qualitative criteria will be applied to extracted FSs items from a corpus. The model mainly consists of three basic phases for identifying the FSs; in each phase the listed items undergo different sorts of analysis until the final refined list of FSs is achieved. This model was developed through a thorough literature analysis.

3.1 Issues of frequency, extent and identification

The main objective of this research is to arrive at a list that is beneficial for pedagogic utility, and frequency tends to be one of most important indicators for this usefulness (e.g. Nation, 2001; O'Keeffe et al., 2007). In English, several single word and FS lists have been used for a long time and have proved to be a fundamental pedagogical tool for designing and developing various kinds of teaching materials and language curriculums (e.g. Nation, 2001; O'Keeffe et al., 2007; Schmitt & Martinez, 2012). Nation & Waring (1997: 18) assert that including high frequency multi-word expressions is as important as the inclusion of single words in the frequency-based lists.

It is reasonable to suppose that this frequency-usefulness relationship also applies to FSs; hence, in this research, frequency is one of the essential statistical criteria for constructing the current list of FSs. Another point about frequency is related to the phrase length. Statistically, there is an inverse connection between the length of phrase and its frequency, so long phrases are always less frequently used in the language. Hence, the focus of this research was on phrases from two words to no longer than four contiguous words.

Regarding the extent of this list, any specific-purpose list must stop at some point to be widely used, so the current study adopted a threshold of 5,000 items. This cut-off point is consistent with many previous frequency-based lists developed for LP (Davies & Gardner, 2010; Milton, 2009; Capel, 2010). The identification of FSs combines quantitative and qualitative approaches, to be of maximum utility. Semantic and linguistic specifications will be considered in the development stages of this list; hence, in the final refined list, phrases which realize meanings or functions are included.

3.2 The corpus source of the language data

We use the Leeds modern standard Arabic Web Corpus (LAWC), which consists of about 176 million words and parts-of-speech (POS), tagged by the *SALMA* (Standard Arabic Language Morphological Analysis) *POS-tagger* (Sawalha & Atwell, 2013). The LAWC corpus is available on the *Sketch Engin*e website (Kilgarriff et al., 2014; https://www.sketchengine.co.uk). It was selected for several reasons,

including its reputation, diversity and size. LAWC was initially collected by Sharoff (2006) for translation studies research and was used in several corpus-based Arabic studies. The corpus is considered to be representative of various written and spoken language genres; it covers various topics, classified into eight main categories (e.g. science, politics, arts and business). LAWC is one of the few publicly available large Arabic corpora; many corpus linguistics studies point out the importance of the corpus size in the overall improvement of the corpus studies results (e.g. Biber et al., 1999; Hunston, 2002; Lee & Cantos, 2002).

3.3 The selection criteria

Based on the working definition of this study, this section presents the selection criteria of FSs, focused on the semantic features of the phrase regardless of its structural varieties. Because of this, we preferred to adopt a semi-automatic extraction approach from the corpus, which combined the use of manual and automatic techniques in the development of the FSs list.

The well-known limitations of fully automatic identification of FSs, especially in a morphologically rich language, and the lack of an Arabic corpus annotated with MWEs justify the use of manual methods for extracting FSs with particular semantic features. The criteria used in this study are based on an intensive analysis of relevant previous English and Arabic research, see Section 2. The criteria were used as a guide for selecting FSs, thus any cluster (word-sequence) which met at least one of these four criteria from the n-gram candidates list should be considered as a potential FS candidate that is subject for further validation. Our selection criteria for FSs identification were:

i. Does the expression, or part of it, lack sematic transparency?

This means that the meaning of the phrase is not purely derived from its component parts, such as *kick the bucket* which means "to die" and in Arabic, *antql ʾilā raḥma Allah* (lit. "passed to the mercy of God") which means *māt* ("to die"). However, when applying this criterion, we should bear in mind the fact that fully semantic transparent phrases are rare in language (Taylor, 2006); therefore, expressions

with any degree of non-compositionality will be taken into consideration in the process of FSs identification.

ii. Does the expression show any sort of linguistic idiomaticity?

This criterion can be applied at all linguistic levels of analysis. For instance, at the lexical level the popular Arabic phrase, *ʿalā arraġm man* (''although'') is an idiomatic expression because the lexical items in this phrase cannot be replaced with any other similar words. In addition, the phrase *raj ʿ bakfī ḥanīn* ("came back with Hunain's shoes") is another example which shows a type of morphosyntactic idiomaticity because it presents resistance to morphological or syntactic transformations.

iii. Is the expression related to a specific situation or register?

This criterion is related to the pragmatic properties of the phrase; that is, in every language, many expressions are strongly attached to specific occasions and usually tend to be used to convey a communicative meaning related to the situation. Examples include *excuse me* and *happy birthday* and in Arabic*, šakrā lak* ("thank you") and*, maʿ assalāma* ("goodbye").

iv. Can the expression be paraphrased or translated into a single word?

This criterion helps to identify a FS; in English several studies have used a parallel translated corpus to detect different kind of MWEs (Nerima et al., 2003; Smadja et al., 1996) by analysing their equivalent in other languages, for instance, the Arabic phrase*, baġḍ annaḍr ʿan* is translated in one equivalent word in English ("regardless'').

3.4 Stages of constructing the FSs list

The FSs extraction was achieved in three main stages. Firstly, the statistical phase: in this phase, co-occurring words were automatically extracted using the *Sketch Engine* n-gram tool with a frequency

threshold of ten per million words, because the focus in this experiment was on the most frequent n-grams. The resulting list of FSs was ranked in descending order by their frequency. Secondly, the qualitative phase aimed to apply the above qualitative criteria to the automatically extracted list; any FSs which met one of the pre-determined qualitative criteria were included in the final list at this stage. In addition, a validation exercise was conducted on a sample to ensure the validity of the selection criteria and demonstrate the possibility of replicating the resulting list by other researchers. Finally, the linguistic classification and annotation phase was undertaken to analyse the listed items linguistically by applying POS tagging and classifying the refined list items into different categories according to their linguistic properties; and to annotate the FS with corpus examples. Figure 2 shows the proposed hybrid model for extracting a pedagogical listing of FSs.

INSERT FIGURE 2 HERE

Figure 2. Diagram of the proposed hybrid model for extracting a list of FSs

3.4.1 *Statistical phase*

As a pre-processing step to the statistical phase, all unnecessary particles and mistyped words were removed from the text extracted from the corpus in *Sketch Engine*. In addition, orthographical normalisation was performed; for instance, the Arabic (آ-إ-أ) Alef-letter variations were normalised to (ا) *A*. This pre-processing phase assisted in reducing the initial list of candidates. Using the n-gram tool of the *Sketch Engine* website, with a frequency threshold of 10 per million words and a cut-off point of four for the MI association score, the search rendered a list of 5,115 n-grams. The defined search span included all n-grams between two and four words. We assigned the minimum frequency to concentrate on the most frequent FSs. Several studies in phraseology provide psycholinguistic evidence that a co-occurrence of words with a minimum MI score of 3 can be considered as a collocation (Church & Hanks, 1990; Stubbs, 1995). The n-grams were then ranked in descending order by frequency, and annotated with the association measure scores.

3.4.2 *Qualitative phase*

This was the most time-consuming stage in developing the ArFSs list, but essential since the objective was to arrive at a meaningful and pedagogically relevant list. The use of computational techniques in

the process of finding qualitative criteria might lead to an inadequate listing. Hence, the study had to rely on manual processing at this stage. Firstly, in the light of the established selection and exclusion criteria, a number of list items were manually removed from the initial listing, as a pre-processing step to the qualitative phase. The following are reasons for removing items from the primitive list of FSs:

i. The phrase involved an abbreviation, a proper noun and numbers;
ii. Dialectical Arabic words or expressions were removed in the light of the fact that this list concentrated only on classical and standard modern ArFSs;
iii. Items appeared on the listing more than once because of variant spellings (their frequency was added to the correct spelling phrase);
iv. The phrases were meaningless such as word sequences that consisted merely of articles or prepositions;
v. Named entity constructions: names are generally not included in general language FS lists;
vi. Redundancy: items appeared in 2- and 3- or 4-gram word sequences with little variation.
vii. Transparency: items for with the meaning was directly derived from their component words, and which did not meet any other inclusion criteria.

Once this pre-processing step had been completed, a list of the most distinctively and meaningful FSs was developed comprising 1,773 statistically ranked items. The main qualitative analysis step involved carefully going through the entire set of n-gram word-sequences, one by one, to apply the qualitative criteria. Any n-gram which met at least one of the pre-defined criteria was included in the ArFSs list. On several occasions it was necessary to consult the corpus concordance and a number of Arabic dictionaries to investigate in depth the meaning of the FSs in various contexts. For instance, the phrase, *wafqā la* ("according to") did not seem to be formulaic at first glance but deeper consideration of its meaning in a different context yielded an idiomatic result. Another phrase, *bamā ʾan* ("so that") seemed to be meaningless, but when consulting the corpus concordance, its meaning became clear. When this process had been finished, a list of statistically and qualitatively filtered n-grams was created involving 608 remaining items of FSs.

After the list items underwent the qualitative filtering, a validation rating exercise was conducted to judge whether the selection of qualitative criteria could be applied by other researchers.

The initial list consisted of more than 5,000 items, which is a very large number for a full validation exercise; therefore, a random sample of 350 n-gram word-sequences was extracted. The first 50 n-grams were used as a training sample, before the conduct of the actual rating exercise on 300 n-grams. Table 1 shows the task set for the validation exercise.

Table 1. The task for the validation exercise

Tick all the phrases that match the qualitative criteria.
Use the corpus concordance tool or/and Arabic dictionaries if you need to understand the meaning of any n-gram items.
If you are hesitant, you can make notes about the hesitation in the comment column.

Nu	FSs	Freq	Yes	No	Comments
1	من خلال	59900			
2	أكثر من	54114			
3	عليه السلام	39907			
4	من اجل أن	37889			
5	بالنسبة لـ	31243			

The independent assessor for this exercise was a PhD researcher in applied linguistics with robust experience of teaching Arabic as a second language and the development of language learning and teaching materials. To ensure the judge's familiarity with FS research, prior to the assessment process, he was introduced to the key research in FSs along with an informative detailed discussion session about the concept and the selection criteria of FSs. The assessor was then presented with a detailed written document that outlined the scope and objectives of the research in general and gave special explanations about the qualitative selection criteria and the process of applying these criteria by the researcher. Then, the assessor was asked to carefully read the whole set of n-grams line by line, to select any n-gram word-sequences that met at least one of the qualitative selection criteria. In case of uncertainty about any list items, the assessor was told to consult the corpus concordance, or Arabic dictionary, and to write a note about his choice.

3.4.3 *Linguistic analysis and classification phase*

This phase mainly concentrated on linguistic analysis of our list items at various levels. Only shallow morphological, structural and semantic analyses were covered in this phase; more advanced linguistics analyses such as syntactic, lexicographic and pragmatic will be considered in upcoming research. Our classifications were consistent with previous studies on Arabic MWEs (e.g. Hawwari et al,. 2014;

Meghawry et al., 2015). In the first step, the automatic *MADAMIRA* POS tagger (Pasha et al., 2014) was applied to all list candidates, to classify them into POS structural pattern categories.

Then we examined the level of ArFSs compositionality. The extracted FSs varied in their degree of idiomaticity, this means that the meaning of the FSs varied in its relation to the phrase component parts; some phrases can be easily understood if you know the meaning of its component parts, while others have a different meaning that is irrelevant to its component parts. Mel'ćuk (1998) presents semantic classifications of phrases with regard to their degree of idiomaticity; the first category is 'full phrasemes', i.e. the meaning of the phrase cannot be derived from its component parts. The second one is 'semi-phrasemes', i.e. the meaning of the phrase matches the meaning of its component parts, but it has an additional meaning which is not related to its component parts. The third category is 'quasi-phrasemes': here, the meaning of the expression is derived directly from one part of the phrase and partially or indirectly derived from the other one. These semantic opacity classifications can be adopted in our study. Hence, the ArFSs in this study were classified into three main categories: full phrasemes, semi-phrasemes, and quasi-phrasemes.

The last step of this stage was related to extracting a corpus-based example of each phrase, which represented the actual use and context of the ArFSs. This step was targeted to enhance the pedagogical applications of this list. For each n-gram, a random list of concordance lines was generated and the researcher very carefully read all the concordance lines to discover the in-depth meaning of each expression and then classified them into categories according to their meanings in different contexts. The most frequent meaning was then added to the FSs list as a good example of each expression.

4. Results and discussion

In the statistical phase, we extracted a set of 5,115 n-grams from our corpus with a frequency above 10 per million words and an MI score above 4. Table 2 shows examples of the high-frequency 2-grams.

Table 2. Sample of the initial unedited 2-grams list derived from LAWC.

Bigram	**Translation**	**Raw Freq**	**Association Measures**	
			MI	**log likelihood**
بشكل *bašk*l	In a manner	68,964	5,182	451,430

من خلال *man ḳalāl*	Through	59,900	4,507	289,756
بسبب basbb	Because	49,771	5,071	329,927
بالنسبة *bālnsba*	For	47,091	5,407	341,640
من أجل *man ʾajl*	In order to	37,889	5,378	263,642
وسائل الإعلام *wasāʾil alʾiʿlām*	The media	59,98	11,250	86,598
سبيل المثال *sabīl almaṯāl*	For example	9,452	12,252	156,386
يوم القيامة *yawm alqayāma*	The Day of Judgement	8,323	10,299	113,011

In the qualitative phase, the 5,115 n-grams were first examined to remove abbreviations, numbers, dialect words, variant spellings, meaningless function-word-only sequences, names, redundant n-gram subsequences, and clearly transparent word-combinations; this manual filtering removed 3,342 candidates, leaving 1,773 n-grams. Table 3 shows examples of the phrases and the reasons why they were deleted from the initial list of n-grams.

Table 3. Examples of excluded FSs, with the reasons for their exclusion

Excluded expression	**Reason**
مش عايز, *mašʿāyz* (“I do not want”)	Dialectical language
الحرب العالمية, al*ḥarb al*ʿ*ālmya* (“World war”)	Proper noun
هذه الفكرة, ha*ḏ*h alfakra (“this idea”)	Transparency
على الرغم من, *ʿalā arraġm man* (“Although”)	Redundancy

Each of the remaining 1,773 n-grams was manually assessed against the four qualitative criteria; 1,165 were judged to not meet any of the four qualitative criteria. The final result was a refined list of 608 ArFSs which met at least one of the pre-determined criteria. The criteria were validated in an inter-annotator agreement exercise. In the training sample, the assessor selected 11 out of 50 items and he only missed two of the items we judged as FSs. However, he pointed out in the comment column that he did not know the exact meaning of these phrases. Hence, he was told to consult the corpus concordance to clarify their meaning. Once he realised their meaning, the assessor decided to include them with the selected items.

After finishing the exercise, the assessor reported that the qualitative selection criteria were very clear that he usually had not taken a long time to decide whether to include or exclude the list items. He also added that the excluding criteria provided were very beneficial in the excluding of many n-gram clusters. The assessor missed only 17 FSs in the test sample of 300 n-grams. The inter-rater

reliability was high, overall the gold standard and assessor were 94.4% in agreement. This is an encouraging inter-rater reliability result, which demonstrated the reproducibility of our resulting list. The high agreement is probably not surprising as the selection and the excluding criteria were carefully and clearly defined.

In the linguistic analysis and classification phase, we enriched the final lexical resource of 608 ArFSs by adding linguistic classification, and corpus examples. Table 4 shows examples of ArFSs with POS tags from the MADAMIRA POS tag-set. Regarding ArFSs structures, the list items appeared to belong to well-known structural patterns of Arabic MWEs. Table 5 shows the structural categories of our ArFSs list. The classification of ArFSs by level of compositionality is illustrated in Table 6. As a practical addition to the resource, each ArFS was annotated with a representative corpus example. Table 7 shows several FSs with their corpus-based examples.

Table 4. FSs examples with their POS Tags

POS	Example
Nouns	نظراً لـ *naḍran la*
Adjectives	متعلقة بـ *matʿlqa ba*
Adverbs	هنا وهناك *hanā wahnāk*
Verbs	يؤدي إلى *yaʾdī ʾilā*
Particles	لابد *lābd*
Prepositions	في إطار الوصول *fī ʾiṭār alwaṣūl*
Conjunctions	وبـالتالي *wb_AltAly*
Interjections	سبحان الله *sbHAn Allh*

Table 5. Examples of Structural patterns of Arabic FSs

Structure	Example	Translation
preposition + noun	بمناسبة, bamnāsba	By the way
noun +preposition	ردًا على, radan ʿalā	In esponse to
preposition + noun+noun	في نهاية المطاف, fī nahāya almaṭāf	Eventually
Interjections + noun	سبحان الله, sabḥān Allah	Glory be to Allah
Conjunctions + preposition + noun	وبالتالي, wabāltālī	Thus
preposition + adjective + noun	في بعض الأحيان, fī baʿḍ alʾaḥyān	Sometimes
Verb + noun	رحمه الله, raḥmh Allah	May Allah have mercy on him
preposition + Pronouns +noun +conjunction + noun	عليه الصلاة والسلام, ʿalīh aṣṣalā wassalām	Peace and mercy be upon him

Adjective + Pronouns	غالبا ما, ġālbā mā	Often
noun +conjunction	رغم أن, raġm ʾan	Though
Preposition +adjective	من الضروري, man aḍḍarūrī	It is necessary
Adjective+ Preposition	مرتبط بـ, martbṭ ba	Linked to
Adverb+ conjunction+ Adverb	هنا و هناك, hanā wahnā	Here and there
Noun + noun	ذات الصلة, ḏāt aṣṣala	Related to
Conjunction+ Particle+ noun	ولا سيما, walāsīmā	In particular

Table 6. Semantic compositionality opacity levels of the FSs

Semantic degree	Example
Full phrasemes	بالطبع, *bālṭbʿ* (“of course”)
Semi-phrasemes	إلى حد ما, *ʾilā ḥad mā* (“to somewhat”)
Quasi-phrasemes	السياسية الخارجية *assayāsa alḳārjya* (“Foreign policy”)

Table 7. Several FSs with their corpus-based examples

FSs	Sentence Example	Translation
على الرغم من, *ʿalā arraġm* man (‘‘although’’)	كان من أبطال الفيلم الأساسيين على الرغم من بعض العثرات في تقليد اللهجة.	He was a major film hero despite some pitfalls in imitating the dialect.
بغض النظر عن, *baġḍ* anna*ḍr ʿan* (‘‘regardless’’)	التعرض ل أشعة الشمس لفترات طويلة خطر يهدد جميع الأعمار بغض النظر عن كون من يتعرض لها رضيع أو طفل أو شاب.	Prolonged exposure to sunlight is a threat to all ages, regardless of whether they are exposed to an infant, a child or a young person.
وبالتالي, *wabāltālī* (“therefore”)	أن تناول طعام صحي وسليم مع اللعب النشيط يؤدي إلى صحة جيدة، وبالتالي فان الصحة الجيدة تؤدي إلى شهية جيدة.	Eating healthy food with active lifstyle leads to good health, therefor this leads ultimitly to good appetite.
على سبيل المثال, *ʿalā sabīl* alma*ṯāl* (‘‘for example”)	يجب أن تكون التكاليف في الإسلام حسب طاقة المكلف فلا يطلب منه على سبيل المثال إلا اثنين ونصف بالمائة من ربحه السنوي الصافي كزكاة.	The duties in Islam must be according to the capacity of the taxpayer, for example, only two and a half percent of his net annual profit is required as a Zakah.

The hybrid model adopted in this research enabled us to take advantage of automatic and manual extraction of FSs that resulted in the development of meaningful and manually validated list items that can be used in different pedagogical and NLP tasks. To facilitate the usability and accessibility for the end-users of this resource, particularly for language learners and teachers, sentence examples were provided for each list item.

In terms of the linguistics processing stage, the overall results of the POS analysis showed that most of the FSs that met the selection criteria were phrases that began with prepositions. Figure 3 illustrates the overall result for ArFSs.

INSERT FIGURE 3 HERE

Figure 3. FSs list distribution by the POS of the head word.

However, this dominance of prepositional phrases in the extracted FSs list might be due to the focus on the most frequently occurring n-grams in our method. The large number of prepositional phrases demonstrates the key role of this kind of phrase in everyday language use, so this initial result indicates the need for a special consideration of prepositional ArFSs in further experiments. Prepositional phrases are very important in NLP and second LP because they are considered as highly ambiguous and difficult in language processing and learning tasks.

The semantic analysis of the FSs enhances the utility of the extracted ArFS list in different practical NLP tasks; for instance, knowing the degree of idiomaticity of the phrase can be of great benefit in increasing the precision and robustness when trying to integrate this list into an NLP system. Knowing the non-compositional FSs enables an NLP system developer to treat them as a single word, which ultimately increases the overall accuracy of NLP system output. In addition, these kinds of FSs are usually considered as the most difficult phrases to learn for non-native speaker/learners of Arabic. Hence, the concentration on this kind of FS in the design of learning and teaching materials will have a positive impact on accelerating the process of second language acquisition.

6. Conclusions

This paper has described a methodical approach to develop a list of ArFSs for use in NLP and LP. To ensure the quality of the list items, the hybrid model was applied in the process of extracting ArFSs from a large corpus. This methodology enabled us to go beyond a solely automated extraction process that would only extract FSs of limited value in NLP and LP. The different levels of analysis that the FSs underwent contributed to the identification of valuable FSs that can be of great benefit for NLP and non-native speaker/learners.

The methodology in our research is not without limitations, including the use of only two expert Arabic linguists in the procedures of applying the qualitative criteria and in the reliance on a raw-text Arabic corpus because of the lack of linguistically annotated Arabic corpora. The present study research is an effort toward more intensive corpus linguistics research on Arabic FSs. In addition, this research presented a model for identifying FSs based on several well-defined FSs criteria which can also be applied to other varieties and languages. The hybrid mixed-methods approach has been demonstrated for Modern Standard Arabic, so a next step will be to apply the same methodology to extract specialised ArFSs lists for other varieties of Arabic, using other Arabic corpora. For example, to build lists of ArFSs specific to Quranic Arabic and Classical Arabic (Alrehaili & Atwell, 2017), using the Quranic Arabic Corpus (Dukes & Atwell, 2012) and the King Saud University Corpus of Classical Arabic (Alrabiah et al., 2014). Other options include to build lists of ArFSs for specific Arabic dialect corpora (Hassan et al., 2013; Alshutayri et al., 2016) and ArFSs for specific genres of Arabic, such as children's Arabic (Al-Sulaiti et al., 2014), second-language learner Arabic (Alfaifi et al., 2014) or Arabic social media (Alshutayri & Atwell, 2017; forthcoming).

The initial list of 5,115 statistically-frequent n-grams was pared down to a subset of 608 "true" ArRFs which met linguistic and semantic criteria. A side-effect of this manual filtering is that we have a list of 4,507 "rejects": candidate n-grams which should NOT be included in an ArFSs list. This can be a useful resource in further research to extract ArFSs in other varieties or sublanguages of Arabic: normally we expect that if a candidate is not a true ArFS in Standard Arabic then it is also not a true ArFS in the sublanguage. This assumption still has to be tested, but if correct, we have a resource to filter out some candidates generated from other sublanguage corpora.

Another use of the "reject list" is as a gold standard for Machine Learning of "true" ArFSs, for example in evaluation contests such as SEMEVAL, the annual Semantic Evaluation contest run by the Association for Computational Linguistics Special Interest Group on the Lexicon (ACL SIGLEX). We can provide lists of "valid" and "reject" ArFSs candidates, as training and evaluation datasets, for a SEMEVAL contest where the task is to classify an Arabic n-gram as a true ArFS (or not).

The present study is a contribution to a research community effort to construct a comprehensive repository of Arabic lexical resources for NLP and LP. Follow-up research will extend the current list to include less frequent ArFSs, and special attention will be paid to the analysis and extraction of prepositional ArFSs. In addition, different experiments can be conducted using a combination of

knowledge-based and data-driven approach to arrive at a more valid and comprehensive result. The research will also aim to integrate the expert and non-native learners' judgements in the selection and classifications of ArFSs to improve the quality and accessibility of this language resource. The ArFSs items can be enhanced with a comprehensive annotation scheme that aims to cover additional linguistic features of ArFSs, including phonological, orthographical, syntactic, semantic and pragmatic features. The final ArFSs repository can be integrated into a free access online e-learning environment to make the most of this significant language resource.

Acknowledgments

This research has been generously supported by Umm Al-Qura University, Mecca, Saudi Arabia. This paper extends our earlier work on ArFSs reported in (Alghamdi et al., 2016), (Alghamdi & Atwell, 2017).

Notes

1. The German standard DIN 31636 is used for rendering Romanized Arabic as described in the Appendix.

References

Abou-Saad, A. (1987). A Dictionary of Arabic Idiomatic Expressions. Beirut: Dar ElIlm Lilmalayin.

Alfaifi, A., Atwell, E. & Hedaya, I. (2014). Arabic Learner Corpus (ALC) v2: A new written and spoken corpus of Arabic learners. In S. Ishikawa (Ed.), Proceedings of Learner Corpus Studies in Asia and the World, (pp. 77-89). Kobe: Kobe University. Retrieved from http://eprints.whiterose.ac.uk/79561/ (last accessed April 2019).

Alghamdi, A., Atwell, E., & Brierley, C. (2016). An empirical study of Arabic formulaic sequence extraction methods. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds), *Proceedings of LREC'2016 Language Resources* and Evaluation Conference (pp. 502-506). Portoroz: LREC. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/pdf/126_Paper.pdf (last accessed April 2019).

Alghamdi, A., & Atwell, E. (2017). نحو معجم حاسوبي للمتالزمات اللفظية في اللغة العربية المعاصرة. [Towards a Computational Lexicon for Arabic Formulaic Sequences]. In Proceedings of TICAM The International Conference on Information and Communication Technologies. Retrieved from http://event.ircam.ma/data/papers2016/18.pdf (last accessed April 2019).

Alrabiah, M., Al-Salman, A., Atwell, E., & Alhelewh, N. (2014). KSUCCA: A key to exploring Arabic historical linguistics. International Journal of Computational Linguistics, 5(2), 27-36.

Alrehaili, S., & Atwell, E. (2017). Extraction of Multi-Word Terms and Complex Terms from the Classical Arabic Text of the Quran. International Journal on Islamic Applications in Computer Science and Technology, 5(3), 15-27.

Alshutayri, A., Atwell, E., Alosaimy, A., Dickins, J., Ingleby, M., & Watson, J. (2016). Arabic Language WEKA-Based Dialect Classifier for Arabic Automatic Speech Recognition Transcripts. In P. Nakov, M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann & S. Malmasi (Eds.) *Proceedings of VarDial'2016* Third Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 204-211. Osaka: COLING. Retrieved from https://aclanthology.info/papers/W16-4826/w16-4826 (last accessed April 2019).

Alshutayri, A., & Atwell, E. (2017). Exploring Twitter as a Source of an Arabic Dialect Corpus. International Journal of Computational Linguistics, 8(2), 37-44.

Alshutayri, A., & Atwell, E. (forthcoming). A social media corpus of Arabic dialect text. In C. Wigham & E. Stemle (Eds.), Computer-Mediated Communication and Social Media Corpora. Clermont-Ferrand: Presses Universitaires Blaise Pascal.

Al-Sulaiti, L., Abbas, N., Brierley, C., Atwell, E., & Alghamdi, A. (2016). Compilation of an Arabic children's corpus. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds), *Proceedings of LREC'2016 Language* Resources and Evaluation Conference (pp. 1808-1812). Portoroz: LREC. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/summaries/142.html (last accessed April 2019).

Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), Advances in Natural Language Processing (pp. 87-98). Berlin: Springer.

Attia, M., Tounsi, L., Pecina, P., van Genabith, J., & Toral, A. (2010). Automatic extraction of Arabic multiword expressions. In É. Laporte, P. Nakov, C. Ramisch, A. Villavicencio (Eds), Proceedings of MWE'2010 Workshop on Multiword Expressions: from Theory to Applications (pp. 19-27). Beijing: COLING. Retrieved from https://www.academia.edu/951288/Automatic_Extraction_of_Arabic_Multiword_Expressions (last accessed April 2019).

Atwell, E. (1982). LOB corpus tagging project manual postedit handbook. Research report, University of Lancaster. Retrieved from https://www.researchgate.net/publication/246707360_LOB_Corpus_tagging_project_post-edit_handbook (last accessed April 2019).

Baldwin, T. (2004). Multiword expressions, an advanced course. Paper presented at The Australasian Language Technology Summer School (ALTSS 2004), Sydney, Australia.

Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In F. Bond, A. Korhonen, D. McCarthy & A. Villavicencio (Eds), Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (pp. 89-96). Association for Computational Linguistics. Retrieved from https://aclanthology.info/papers/W03-1812/w03-1812 (last accessed April 2019).

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau, Handbook of Natural Language Processing (2nd ed., pp. 267-292). Boca Raton, FL: Chapman and Hall/CRC.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman Grammar of Spoken and Written English. Harlow: Longman.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language Teaching Research, 10(3), 245-261.

Butt, M. (1999). A Grammar Writer's Cookbook. Stanford, CA: CSLI.

Calzolari, N., Lenci, A., & Quochi, V. (2002). Towards multiword and multilingual lexicons between theory and practice. Paper presented at LP2002, Urayasu, Japan.

Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. English Profile Journal, 1, e3. https://doi.org/10.1017/S2041536210000048

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22-29.

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. Journal of Pragmatics, 3(3-4), 239-266.

Coxhead, A. (2000). A new academic wordlist. TESOL Quarterly, 34(2), 213-238.

Davies, M., & Gardner, D. (2013). A Frequency Dictionary of American English: Word Sketches, Collocates and Thematic Lists. Abingdon: Routledge.

Dawood, M. (2003). A Dictionary of Arabic Contemporary Idioms. Cairo: Dar Ghareeb.

Dipper, Stefanie. (2003). Implementing and Documenting Large-Scale Grammars – German LFG (Unpublished doctoral dissertation). Institut für maschinelle Sprachverarbeitung der Universität Stuttgart, Germany.

Dorgeloh, H., & Wanner, A. (2009). Formulaic argumentation in scientific discourse. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. Wheatley (Eds.), Formulaic Language Volume 2. Acquisition, Loss, Psychological Reality, and Functional Explanations (pp. 523-544). Amsterdam/Philadelphia, PA: John Benjamins.

Dukes, K., & Atwell, E. (2012). LAMP: A multimodal web platform for collaborative linguistic analysis. In N.Calzolari, K. Choukri, T. Declerck, M. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds), Proceedings of *LREC'201*2 Language Resources and Evaluation Conference (pp. 3268-3275). Istanbul: LREC. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/646_Paper.pdf (last accessed April 2019).

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. English for Specific Purposes, 28(3), 157-169.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. Text, 20(1), 29-62.

Fayed, W. K. (2007). A Dictionary of Arabic Contemporary Idioms. Cairo: Abu Elhoul.

Fellbaum, C. (1998). WordNet. Cambridge: MIT Press.

Firth, J. R. (1957). Papers in Linguistics 1934–1951. London: Oxford University Press.

Gralinski, F., Savary, A., Czerepowicka, M., & Makowiecki, F. (2010). Computational lexicography of multi-word units: How efficient van it be? In É. Laporte, P. Nakov, C. Ramisch, A. Villavicencio (Eds), Proceedings of MWE'2010 Workshop on Multiword Expressions: From Theory to Applications (pp.

19-27). Beijing: COLING. Retrieved from https://www.aclweb.org/anthology/W10-3702 (last accessed April 2019).

Guenthner, F., & Blanco, X. (2004). Multi-lexemic Expressions: An Overview. In C. Leclère, É.Laporte, M. Piot & M. Silberztein (Eds.), Lexique, Syntaxe et Lexique-Grammaire [Lexis, Syntax and Lexico-Grammar] (pp. 239–252). Amsterdam/Philadelphia PA: John Benjamins.

Habash, N., & Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In K. Knight, H. T. Ng & K. Oflazer (Eds.), Proceedings of the Conference of American Association for Computational Linguistics (pp. 578-580). Ann Arbor: ACL. Retrieved from https://aclanthology.info/papers/P05-1071/p05-1071 (last accessed April 2019).

Hassan, H., Daud, N., & Atwell, E. (2013). Connectives in the World Wide Web Arabic corpus. World Applied Sciences Journal (Special Issue of Studies in Language Teaching and Learning), 21, 67-72

Hawwari, A., Attia, M., & Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal Arabic. In N. Habash & S. Vogel (Eds.), Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 48–56). Retrieved from https://aclanthology.info/papers/W14-3606/w14-3606 (last accessed April 2019).

Hawwari, A., Bar, K., & Diab, M. (2012). Building an Arabic multiword expressions repository. Paper presented at the ACL 2012 joint workshop on statistical parsing and semantic processing of morphologically rich languages, Jeju.

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4-21.

Isabelli, C. A. (2004). Formulaic Language and the Lexicon. Language Problems & Language Planning, 28(1), 95-98.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1(1), 7–36.

Kjellmer, G. (1990). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), English Corpus Linguistics: Studies in Honour of Jan Svartvik (pp. 111-127). London: Longman.

Lee, D. (2002). Notes to accompany the BNC WORLD edition (bibliographical) index. Retrieved from http://martinweisser.org/corpora_site/BNCWIndexNotes.pdf (last accessed April 2019).

Leech, G. N., Rayson, P., & Wilson, A. (2001). Word Frequencies in Written and Spoken English: Based on The British National Corpus. Harlow: Longman.

Leech, G., Garside, R., & Atwell, E. S. (1983). The automatic grammatical tagging of the LOB corpus. ICAME Journal, 7, 13-33.

Li, W., Zhang, X., Niu, C., Jiang, Y., & Srihari, R. (2003, July). An expert lexicon approach to identifying English phrasal verbs. In E. Hinrichs & D. Roth (Eds.), Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 513-520). Sapporo: Association for Computational Linguistics. Retrieved from https://aclanthology.info/papers/P03-1065/p03-1065 (last accessed April 2019).

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. English for Specific Purposes, 31(1), 25-35.

Martinez, R. (2011). The Development of a Corpus-informed List of Formulaic Sequences for Language Pedagogy (Unpublished doctoral dissertation). University of Nottingham, Nottingham.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. TESOL Quarterly, 45(2), 267-290.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. Applied Linguistics, 33(3), 299-320.

Meghawry, S., Elkorany, A., Salah, A., & Elghazaly, T. (2015). Semantic extraction of Arabic multiword expressions. Computer Science & Information Technology, 5(2), 21-31.

Mel'ćuk, I. (1998). Collocations and lexical functions. In A. Cowie (Ed.), Phraseology. Theory, Analysis, and Applications (pp. 23-53). Oxford: Clarendon Press.

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. Bristol: Multilingual Matters.

Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. Vocabulary: Description, Acquisition and Pedagogy, 14, 6-19.

Nerima, L., Seretan, V., & Wehrli, E. (2003). Creating a multilingual collocation dictionary from large text corpora. In A. Copestake & J. Hajic (Eds.), Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2. (pp. 131-134). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://aclweb.org/anthology/E03-1022 (last accessed April 2019).

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. Wheatley (Eds.), Formulaic Language Volume 2. Acquisition, Loss, Psychological Reality, and Functional Explanations (pp. 387-404). Amsterdam/Philadelphia, PA: John Benjamins.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). From Corpus to Classroom: Language Use and Language Teaching. Cambridge: Cambridge University Press.

Omar, A. (2007). Arabic Multi-word Expressions and Language Resources. Tunis: National Publishing Complex.

Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R.(2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of LREC'2014 Ninth International* Conference on Language Resources and Evaluation (pp. 1094–1101). Reykjavic: LREC. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf (last accessed April 2019).

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), Language and Communication (pp. 191-227). London: Longman.

Peters, A. M. (1983). The Units of Language Acquisition. Cambridge: Cambridge University Press.

Ramisch, C. (2015). State of the art in MWE processing. In C. Ramisch (Ed.), Multiword Expressions Acquisition (pp. 53-102). Berlin: Springer.

Ramisch, C., De Araujo, V., & Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. In J. Cheung, J. Hatori, C. Henriquez & A. Irvine (Eds.), Proceedings of ACL 2012 Student Research Workshop (pp. 1-6). Jeju: Association for Computational Linguistics.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Proceedings of CICLing'2002 Compu*tational Linguistics and Intelligent Text Processing (pp1-15). Berlin: Springer.

Sawalha, M., & Atwell, E. (2013). Accelerating the processing of large corpora: Using grid computing for lemmatizing the 176 million words Arabic internet corpus. In E. Atwell (Ed.), Proceedings of WACL-

2 - 2nd Workshop of Arabic Corpus Linguistics. Lancaster: Lancaster University. Retrieved from http://eprints.whiterose.ac.uk/81622/ (last accessed April 2019).

Schmitt, N. (2004). Formulaic Sequences: Acquisition, Processing, and Use. Amsterdam/Philadelphia, PA: John Benjamins.

Schmitt, N. (2010). Researching Vocabulary: A Vocabulary Research Manual. Basingstoke: Palgrave Macmillan.

Schmitt, N., & Martinez, R. (2012). A Phrasal Expressions List. Applied Linguistics, 33(3), 299-320.

Schneider, N. (2014). Lexical Semantic Analysis in Natural Language Text (Unpublished doctoral dissertation). University of Melbourne, Melbourne.

Scott, M. (2016). WordSmith Tools (Version 6) [Computer software]. Stroud: Lexical Analysis Software.

Seeny, M., Mokhtar, A., & Sayyed, A. (1996). A contextual Dictionary of Idioms [*almu'jm alsyaqi lelta'birat* alastlahiah]. Beirut: Librairie du Liban Publishers.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), WaCky Working papers on the Web as Corpus (pp. 63-98). Bologna: GEDIT.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. Second Language Research, 27(2), 251-272.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 22(1), 1-38.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 2(1), 23-55.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), Formulaic Sequences: Acquisition, Processing and Use, (pp. 153–172). Amsterdam/Philadelphia, PA: John Benjamins.

Wood, D. (2010). Formulaic Language and Second Language Speech Fluency: Background, Evidence, and Classroom Applications. London/New York: Continuum.

Wood, D. (2015). Fundamentals of Formulaic Language: An Introduction. London: Bloomsbury Academic.

Wray, A. (2002). Formulaic language in computer-supported communication: Theory meets reality. Language Awareness, 11(2), 114-131.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), Formulaic Sequences: Acquisition, Processing and Use (pp. 249-268). Amsterdam/Philadelphia, PA: John Benjamins.

Wray, A. (2005). Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Wray, A. (2008). Formulaic Language: Pushing the Boundaries. Oxford: Oxford University Press.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. Wheatley (Eds.), Formulaic Language Volume 1. Distribution and Historical Change (pp. 27–51). Amsterdam/Philadelphia, PA: John Benjamins.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. Annual Review of Applied Linguistics, 32, 231-254.

Wray, A. (2013). Formulaic language. Language Teaching, 46(3), 316-334.

Wulff, S., Swales, J. M., & Keller, K. (2009). "We have about seven minutes for questions": The discussion sessions from a specialized conference. English for Specific Purposes, 28(2), 79-92.

Appendix: The German standard DIN 31636 for rendering romanized Arabic

Original Arabic letter	DIN 31635
أ	ʾ
ب	b
ت	t
ث	ṯ
ج	ǧ
ح	ḥ
خ	ḫ
د	d
ذ	d
ر	r
ز	z
س	s
ش	š
ص	ṣ
ض	ḍ
ط	ṭ
ظ	ẓ
ع	ʿ
غ	ġ
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
هـ	h
و	w
ي	y
◌َ (short vowel)	a
◌ُ (short vowel)	u
◌ِ (short vowel)	i
ا (long vowel)	ā
و (long vowel)	ū
ي (long vowel)	ī

Address for correspondence

Ayman Alghamdi
Arabic Language Institute
Umm Al-Qura University
Saudi Arabia
aamansoori@uqu.edu.sa

Co-author information
Eric Atwell
School of Computing
University of Leeds
e.s.atwell@leeds.ac.uk

Figure 1. Tip of the iceberg shows the complexity of phrases related to the Arabic word *'ayn*

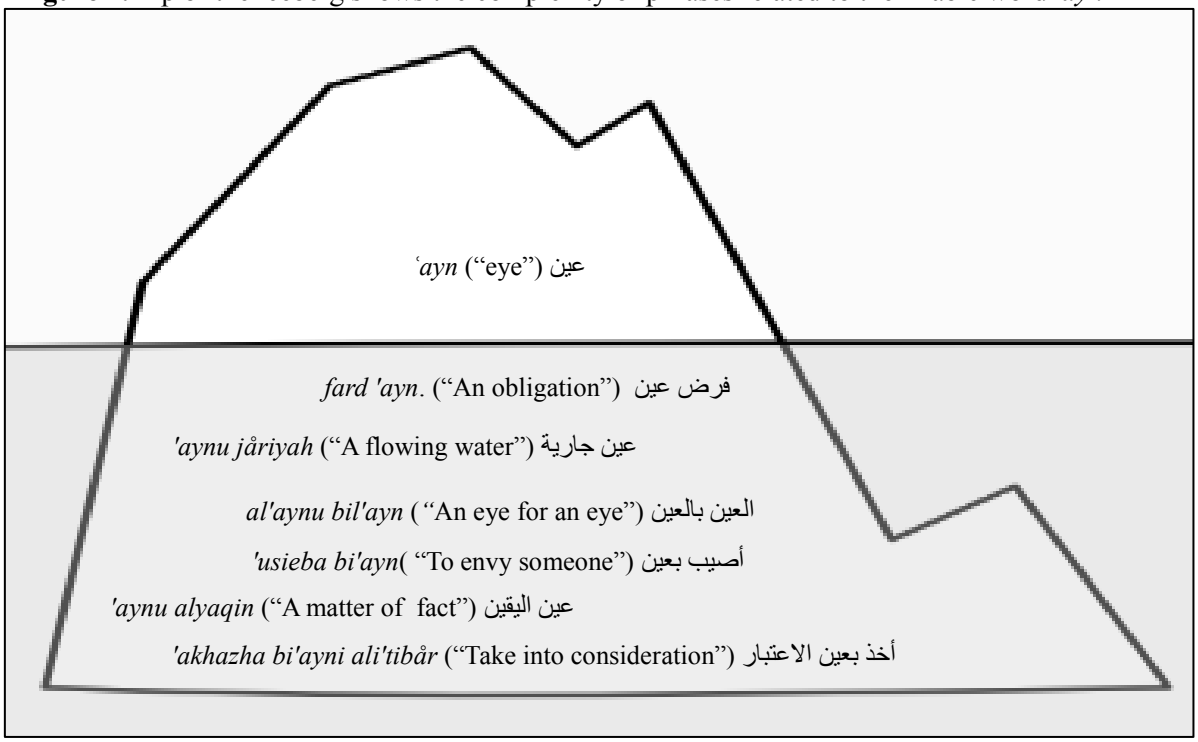

Figure 2. Diagram of the proposed hybrid model for extracting a list of FSs

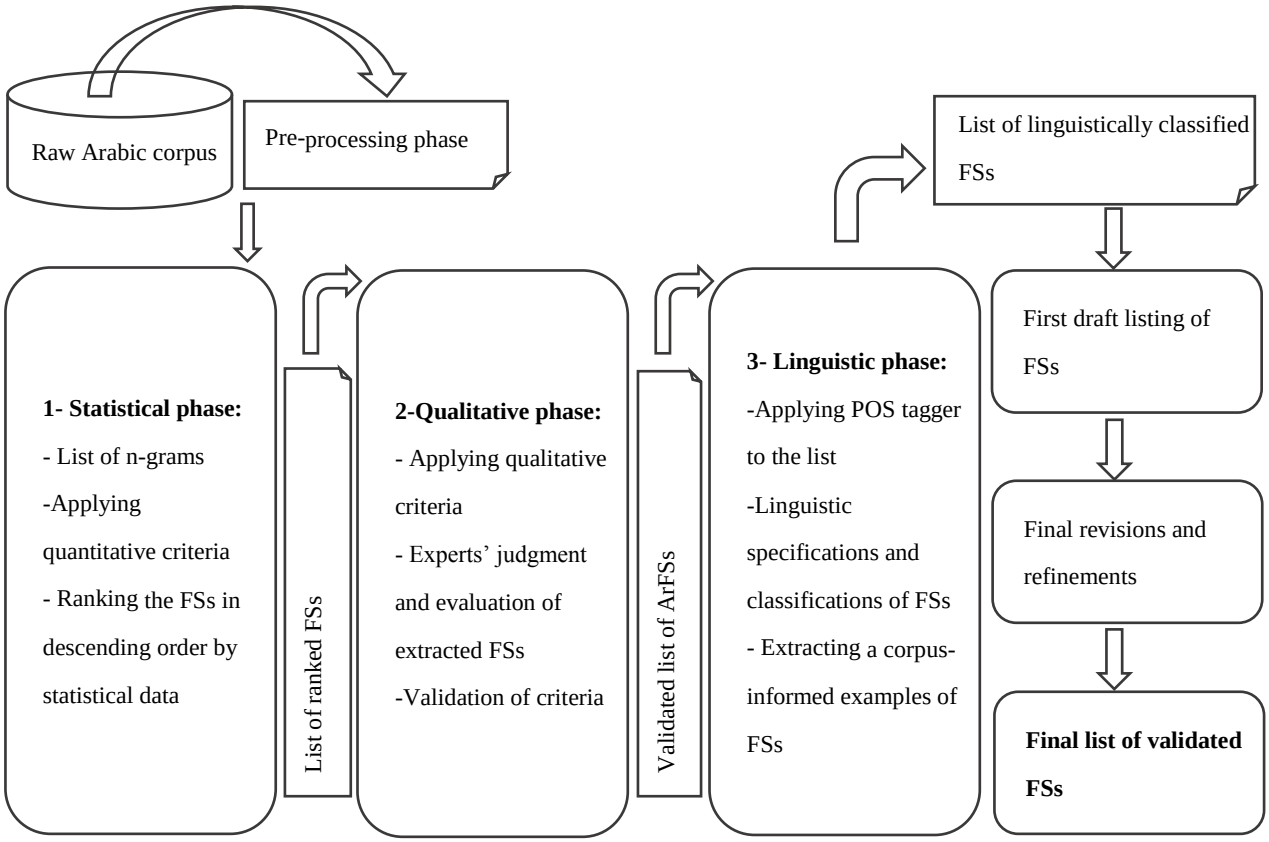

Figure 3. FSs list distribution by the POS of the head word