Introduction
The phenomenon of formulaic sequences or multi-word expressions (MWEs) in languages has attracted the attention of researchers in various language-related disciplines (e.g., linguistics, psychology, language pedagogy and natural language processing ‘NLP’). Hence, this phenomenon has been researched from a number of different scientific angles. A considerable amount of research has emphasised the major role of MWEs in the process of analysing and understanding languages. From the applied 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 words, but is constructed from larger units of meaning, often in the form of phrases or fixed expressions (Fellbaum, 1998). Li et al. (2003) also stated that phrasal verbs constitute approximately one third of the English vocabulary. However, in comparison to English, Arabic MWE research is still at an early stage. Therefore, the role of formulaic language and MWE resources in language pedagogy and NLP and the lack of free access to Arabic MWE lexical resources, justify the conduct of this research to contribute to the remedying of this deficiency by constructing an Arabic corpus-informed FS lexicon for language pedagogy and technology.

Research Significance
The importance of this research is due to a set of factors related to the vital role of integrating the formulaic language knowledge in NLP and language pedagogy. The ignorance of handling MWEs in any language-related tasks will have a negative impact on their final output quality. This is due to the fact that 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 stated that phrasal verbs constitute approximately one third of the English verb vocabulary. However, this large portion of MWEs emphasises their key role in the development of language-related applications. Figure 2 shows an example of the differences between machine translation output before and after integrating English MWE knowledge. Formulaic language research provides evidence that the most frequently used words in languages are not merely the tip of the iceberg of expression-based collocations (e.g., Sinclair, 1987; Martinez & Murphy, 2011). Figure 1 shows the underlying complexity of phrases related to the Arabic word 

Methodology
This research aims to adopt a comprehensive hybrid approach for ArFSs extraction; this will be based on the integration of frequency-based and phraseological approaches, and the combination of knowledge-based and data-driven approaches to identifying ArFSs.

Expected Results
The research results are estimated to achieve the following research objectives:
- To develop a comprehensive computational corpus-informed ArFSs lexicon, which can be incorporated into various Arabic NLP applications.
- To establish standards for describing and encoding ArFS lexical entries at different linguistic levels (morphological, syntactic, lexical and semantic).
- To propose an overall model for ArFSs identification and extraction that will best suit the main objectives of this research.

Research Implications
The pedagogical implications of this lexicon are estimated to facilitate the inclusion of ArFSs in the process of learning and teaching Arabic, particularly for non-native speakers. The computational implications are related to the key role of the ArFSs, as a novel lexical resource, in the improvement of various Arabic NLP tasks and applications. The final novel ArFS lexicon can be integrated into a free access online e-language learning environment to make the most out of it.