An Arabic Formulaic Sequences (ArFSs) Lexicon for Language Pedagogy and Technology

Ayman Alghamdi scaaa@leeds.ac.uk
Eric Atwell e.s.atwell@leeds.ac.uk
Claire Brierley c.brierley@leeds.ac.uk
UCREL
UNIVERSITY OF LEEDS

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 orthographic words, but rather it incorporates longer formulaic sequences (e.g., Pawley & Syder, 1983; Wray, 2002; Nesselhauf, 2005). Other researchers have attempted to develop different 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 the computational perspective, MWEs play a vital role in NLP and many researchers have attempted to construct various types of MWEs repositories in order to apply them in the development of various NLP applications, which include, but are not limited to, MWE identification and extraction, language parsing, information retrieval and machine translation, for example.

Motivation

The vast majority of research in this area has been applied to the English language because of the rich availability of free access machine readable language resources. Recently, Arabic is another language that has received more attention from researchers from different, albeit related, disciplines. However, in comparison to English, Arabic MWE research is still at an early stage. Therefore, the key 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 FSs 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 illustrates the differences between machine translation output before and after integrating English MWEs knowledge. Formulaic language research provides evidence that the most frequently used words in languages are only the tip of expressional icebergs (e.g., Sinclair, 1987; Martinez & Murphy, 2011). Figure 1 shows the underlying complexity of phrases related to the Arabic word عين *ʿayn* 'Eye'.

Thank you for taking the time to read my poster, any comments, feedback or suggestions on this research would be highly welcomed and appreciated.
www.aymanalghamdi.com

Figure 1: Tips of phraseological iceberg shows the underlying complexity of phrases related to the Arabic word عين

Machine Translation Output

Before

This is a delicious hot dog.

هذا كلب ساخن لذيذ .
haḍā kalb sāḵn laḏīḏ
Translated as: 'This is a delicious hot dog'(An animal).

After

This is a delicious hot dog.

هذه شطيرة سجق لذيذة .
haḍh šaṭīra sajq laḏīḏa
Translated as: 'This is a delicious hot dog'(A food 'sausage sandwich')

Figure 2: Machine translation example before and after integrating English formulaic sequences knowledge

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.

Therefore, the selection of formulaic sequences in this lexicon will be based on several pedagogically and NLP relevant criteria from the perspective of second language comprehension and computational linguistics. The research will take advantage of the large, available and free access Arabic corpora that includes written and spoken Modern Standard Arabic Figure 3 shows the adopted model for ArFSs extraction in the first research experiment.

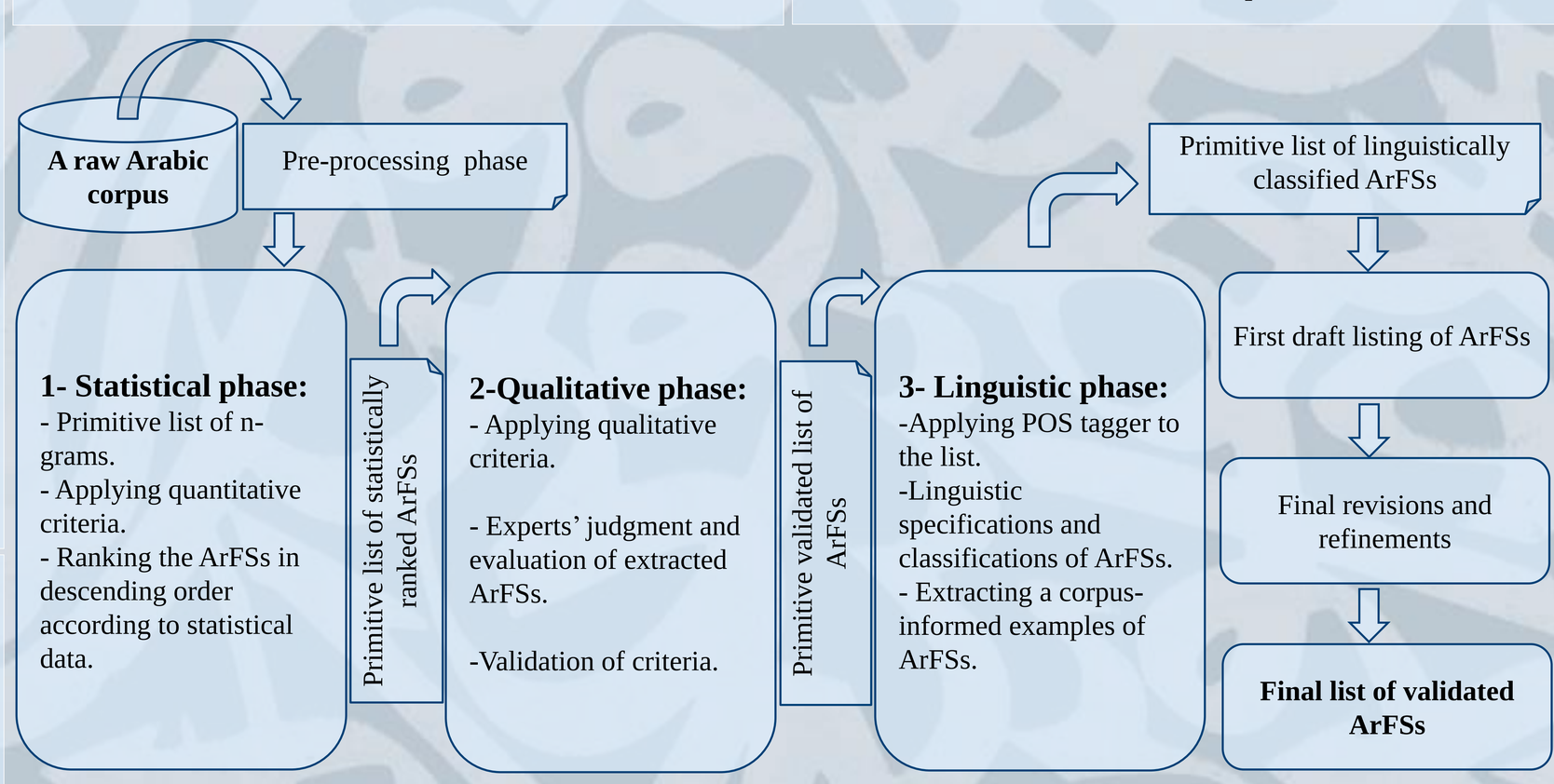

Figure 3: Diagram of the proposed hybrid framework for extracting ArFSs items

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 ArFSs 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 ArFSs lexicon can be integrated into a free access online e-language learning environment to make the most out of it.

References:
Pawley, A., & Syder, F. H. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), Language and communication (pp. 191-226). New York: Longman.
Wray, A. 2002. Formulaic language and the lexicon. Cambridge University Press Cambridge.
Martinez, R. and Murphy, V. 2011. Effect of frequency and idiomaticity on second language reading comprehension. TESOL Quarterly. 45(2), pp.267-290.
Li, W. et al. 2003. An expert lexicon approach to identifying English phrasal verbs. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1: Association for Computational Linguistics, pp.513-520.
Sinclair, J. 1987. Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary. London: Collins ELT.
Fellbaum, C. 1998. WordNet. Blackwell Publishing Ltd.
Nesselhauf, N. 2005. Collocations in a learner corpus. Amsterdam: John Benjamins.