UNIVERSITY OF LEEDS

A COMPUTATIONAL LEXICON AND REPRESENTATIONAL MODEL FOR ARABIC MULTIWORD EXPRESSIONS

Ayman Ahmad O. Alghamdi

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

October, 2018

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

A MEMORABLE DEDICATION

I dedicate this thesis to my beloved father

Ahmad Othman Alghamdi

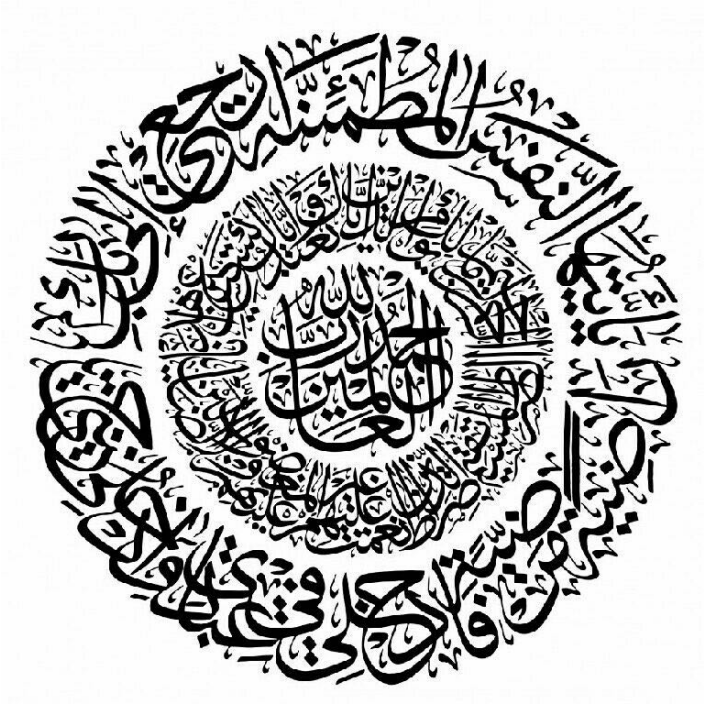
who passed away on Saturday, the 17th of September, 2016.

My dear father, the words are nothing when it comes to showing my feelings towards you,

To you, I am forever grateful…

You are in my mind all the time, and I am missing you every second of life…

Peace and mercy be from ʾallāh forever upon you …



"O reassured soul, return to your Lord, well-pleased and pleasing [to Him], and enter among My [righteous] servants, and enter My Paradise." The Holy Quran, Al-fajr Ch., verse (89:27-30).

PUBLICATIONS

Chapters 4,5 and 7 of this thesis are based on jointly-authored publications. The candidate is the first author of all original contributions presented in these papers: the co-authors acted in an advisory capacity, providing feedback, general guidance, and comments. The work in these chapters of the thesis has appeared in the following publications:

Chapter 4:

Alghamdi, A. 2015. The development of an Arabic corpus-informed list of formulaic sequences for language pedagogy. In: The eighth international Corpus Linguistics conference., University of Lancaster, UK.

Alghamdi, A. and Atwell, E. forthcoming. Constructing a corpus-informed Listing of Arabic formulaic sequences for language pedagogy and technology. Paper submitted to the International Journal of Corpus Linguistics.

Alghamdi, A. and Atwell, E. 2018b. An Arabic corpus-informed list of MWEs for language pedagogy. In: O. L. Dong, J. Lin, W. Xiao, M. Geraldine and P.-P. Pascual, eds. TALC 2018 13th Teaching and Language Corpora Conference. Cambridge, pp. 38–41.

Chapter 5:

Alghamdi, A. and Atwell, E. 2016a. An empirical study of Arabic formulaic sequence extraction methods. In: LREC' 2016 10th Language Resources and Evaluation Conference. Portorož, Slovenia.

Chapter 7:

Alghamdi, A. and Atwell, E. 2017. Towards Comprehensive Computational Representations of Arabic Multiword Expressions In: R. Mitkov, ed. Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, November 13-14, 2017, Proceedings. London: Springer International Publishing, pp. 415–431.

ACKNOWLEDGEMENTS

All the praises and thanks be to our Lord (Allāh), for the blessing and the patience he has provided me to complete this long-term project. It is my duty here to record my sincere gratitude to all the people who supported me during this research journey. First and foremost, this work would not have been possible without the continuous encouragement, insightful ideas, and support of my supervisor Prof. Atwell who guided me throughout all stages of this research; I learnt a lot from you during this journey, and your rich knowledge and academic expertise will always inspire me.

I owe my deepest gratitude to my mother who brought me into this life and always pushed me to achieve my goals; your encouragement played a critical role in keeping me motivated and focused in my work. Special thanks are also due to my family who have made my time here in the UK and Leeds enjoyable and unforgettable; without your continual support and encouragement, this work would not be possible. I am also grateful to all my brothers and sisters who were always concerned about the progress of my PhD, your sincere prayers have helped me overcome many obstacles during my stay here in Leeds.

Special thanks are due to my office mates and all members of the ANLP group at the University of Leeds. I enjoyed my time with you, and we have had informative discussions and valuable suggestions that will pave the way for many research ideas to come. Last, but not least, I owe a great debt of gratitude to Umm al-Qura university for providing me with this scholarship which has opened doors for me to a new bright future. I acknowledge all the people and academic researchers I have met during my PhD journey; your fruitful ideas, comments and informative discussions have always inspired me to achieve my goals.

ABSTRACT

The phenomenon of multiword expressions (MWEs) is increasingly recognised as a serious and challenging issue that has attracted the attention of researchers in various language-related disciplines. Research in these many areas has emphasised the primary role of MWEs in the process of analysing and understanding language, particularly in the computational treatment of natural languages. Ignoring MWE knowledge in any NLP system reduces the possibility of achieving high precision outputs. However, despite the enormous wealth of MWE research and language resources available for English and some other languages, research on Arabic MWEs (AMWEs) still faces multiple challenges, particularly in key computational tasks such as extraction, identification, evaluation, language resource building, and lexical representations.

This research aims to remedy this deficiency by extending knowledge of AMWEs and making noteworthy contributions to the existing literature in three related research areas on the way towards building a computational lexicon of AMWEs. First, this study develops a general understanding of AMWEs by establishing a detailed conceptual framework that includes a description of an adopted AMWE concept and its distinctive properties at multiple linguistic levels. Second, in the use of AMWE extraction and discovery tasks, the study employs a hybrid approach that combines knowledge-based and data-driven computational methods for discovering multiple types of AMWEs. Third, this thesis presents a representative system for AMWEs which consists of multilayer encoding of extensive linguistic descriptions.

This project also paves the way for further in-depth AMWE-aware studies in NLP and linguistics to gain new insights into this complicated phenomenon in standard Arabic. The implications of this research are related to the vital role of the AMWE lexicon, as a new lexical resource, in the improvement of various ANLP tasks and the potential opportunities this lexicon provides for linguists to analyse and explore AMWE phenomena.

The tip of an iceberg shows the complexity of AMWEs related to the word ‘ayn, عين ‘eye’*.

* This image was the winner of the 2018 edition of the Images of Research Competition at the School of Computing, University of Leeds.

‘We [our almighty lord] will, put over every possessor of knowledge is one [more] knowing’. The holy Quran Joseph Ch., verse (12:76).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Meaning	Page
FS	Formulaic Sequence.	5
MWE	Multiword Expression.	1
AMWE	Arabic Multiword Expression.	2
AMWEL	Arabic Multiword Expression Lexicon.	129
NLP	Natural Language Processing.	1
ANLP	Arabic Natural Language Processing.	2
CL	Computational Linguistics.	2
LP	Language Pedagogy.	1
AMs	Association Measures.	29
SA	Standard Arabic.	15
SAP	Stanford Arabic Parser.	125
MA	Mada Amira Morphological Toolkit.	125
APT	Arabic Penn Treebank.	83
BAMA	Buckwalter Arabic Morphological Analyser.	128
LR	Language Resources.	1
MT	Machine Translation.	2
CA	Classical Arabic.	5
LFG	Lexical Functional Grammar.	29
LMF	Lexical Mark-Up Framework.	49
NE	Named Entity.	2
MWT	Multiword Term.	59
ARF	Average Reduced Frequency.	141
MAP	Mean Average Precision.	151
XML	Extensible Markup Language.	44

1 Introduction and Motivation

1.1 Introduction

Multiword expressions (MWEs) are an indispensable part of natural languages and present enormous challenges at different levels of linguistic and computational analysis. This complex phenomenon has attracted the attention of researchers from various scientific backgrounds who have contributed towards increasing understanding and tackling several research challenges encompassing MWE from various perspectives (e.g., linguistics, psychology, language pedagogy (LP), and natural language processing (NLP)).

A considerable amount of research has emphasised the primary role played by MWEs in analysing and understanding human languages. For instance, in linguistics, several theories have been proposed to delineate general descriptions and construct a framework to demonstrate MWE characteristics and behaviour at all linguistic levels (e.g., Mel'čuk, 1998; Gries, 2008; Ruppenhofer et al., 2016; Schneider, 2014; Bejoint, 2013).

In applied linguistic and language pedagogy (LP), researchers have emphasised the crucial importance of including formulaic language and MWEs in the process of second language learning and teaching and learning activities (e.g., Kremmel et al., 2015; Granger and Meunier, 2008; Mel'cuk, 1995). Other research in these areas has attempted to develop different MWE lists or language resources (LRs) that can be used as tools to improve the progress of second language learning in various forms, such as material design, curriculum development, and language testing (e.g., Schmitt and Martinez, 2012; Giacomini, 2017; Gardner and Davies, 2014).

Research in psycholinguistics has emphasised the notion that single orthographic words alone do not constitute our mental lexicon; instead longer lexical units are incorporated through a lengthy and incremental language acquisition process (e.g., Pawley and Syder, 1983; Sinclair, 1987; Wray, 2002; Nesselhauf, 2005).

From an NLP and computational perspective, research has emphasised the importance of integrating MWE knowledge into the improvement of most NLP tasks. Most MWE

research in computational linguistics (CL) and NLP has focused on four research areas. First, building different types of MWE language knowledge bases (LKBs) (e.g., Brooke et al., 2015; Attia et al., 2005; Hatier et al., 2016; Zaninello and Nissim, 2010). Second, finding various computational models for MWE extraction and identification (Pal et al., 2013b; Pecina, 2008; Ramisch, 2015a). Third, proposing and implementing several representational models for formalising MWE knowledge in machine-readable forms (Grégoire, 2009; Odijk, 2013b; Calzolari et al., 2002). Fourth, MWE research related to application-oriented studies has aimed to discover and evaluate different methods for embedding MWE knowledge in the development of various NLP applications, including machine translation (MT), language parsing (LP), information retrieval, semantic search, and named entity recognition (e.g., Carpuat and Diab, 2010a; Luong et al., 2015; Attia, 2006a).

Moreover, most MWE research has primarily been applied to the English language due to the widespread availability of free access language resources and tools, and the interest an extensive international research community has in studying English as the language of science and the most widely spoken language worldwide. However, Arabic has recently received substantial attention from researchers from different, albeit related, disciplines. However, in comparison to English, and despite the current and widespread use of Arabic, MWE research is still at an early stage. Therefore, MWE has a critical role to play in understanding human languages and in the improvement of several ANLP tasks. The lack of research on AMWE, and the need to address the research problems of this thesis, justify the building of a computational lexicon and representational system of Arabic MWEs for language technology.

This chapter presents the motivations that underpin this thesis and the significance and contributions of the research. This will be followed by a brief definition of the research tasks and questions. It will conclude with a brief description of the thesis chapters and related published works.

1.2 Research motivations and significance

Regularity is a typical characteristic of natural languages and can be found at various levels of linguistic analysis. For instance, at the word level in English, it is an easy task for language learners to learn the morphological rule that, to make the verb in the past tense, one should merely add the suffix '*ed*', which means they will be able to

acquire most vocabulary effortlessly. At phrase or sentence levels, the regular semantic rule is that the meaning of a phrase is generally derived from the meaning of its parts, thus when people know the meaning of the words *'blue'* and *'pen'* it is straightforward to predict the meaning of the phrase *'blue pen'*. Unfortunately, this is not always the case, as shown when an attempt is made to extract the meaning of the phrase *'piece of cake'* or *'hot potato'* from the sentence, *'MWE is not a piece of cake topic but it is one of the most important hot potato issues in NLP'*. This is because, in this sentence, there is a violation of the regular rule of compositionality. Thus the meaning can only be derived from the phrase as a semantic whole.

Similar examples can be found in Arabic. For instance, at the word level, the simplest morphological rule for changing words from single to plural forms is merely to add one of these suffixes to the words (ـات - ـين - ـون). However, this is not always operative, as shown in the so-called broken plural in this example (singular: رجل, *rajul*[1]*, man* - plural: رجال, *rijāl* men). At the sentence level, many examples that violate the rule of semantic compositionality can be found; for instance, the meaning of the popular MWE وقع في حيص بيص *waqaʿ fī ḥayṣ bayṣ 'he was in a confused state*' cannot be extracted merely from its individual components because these have little to do with the meaning of this phrase.

However, in-depth corpus-based analyses of natural languages show that such irregularity phenomena are not marginal or trivial issues as human language tends to be more complicated than one might initially think. Most of these complexities are due to the irregular and unproductive nature of language behaviour at various levels of linguistic analysis. This yields several linguistic idiosyncratic phenomena that have exercised the minds of many language learners, linguists and other interested researchers.

This section discuses several issues that constitute the motivations underpinning this thesis. The significance of MWE is illustrated with special attention being paid to AMWE and the prime role of building AMWE LR with comprehensive computational

[1] In the literature, there are several possible transliteration systems for Arabic script. For consistency, in this thesis the German standard DIN 31636 is used for rendering Romanised Arabic, as described in Appendix A. However, readers should be aware that they might encounter various transliterations in the relevant literature.

formalism in the improvement of most NLP tasks. The following subsections briefly address the question: Why do the AMWE research problems tackled in this project matter?

1.2.1 MWE is not a marginal feature of natural languages

MWE constitutes a significant portion of most modern languages and is usually governed by irregular linguistic rules that require close attention and consideration at various levels of processing. Research on various languages has presented evidence to support this claim. In English, most MWE research has been conducted through several corpus-driven studies that confirm the frequency of these types of phrases; they give different estimations of the proportion of these phrases in English, which range from around 30% (Biber et al., 1999) to more than 50% (Erman and Warren, 2000) in spoken and written discourse. Hence, ignoring this significant portion of the language will have a negative impact in any language-related applications. In English WordNet 1.7 (Miller et al., 1990), MWEs constitute 41% of lexical entries, while Li et al. (2003) found that phrasal verbs constitute approximately one-third of the English verb vocabulary. Baldwin and Kim (2010) state that 'the number of MWEs is estimated to be of the same order of magnitude as the number of simple words in a speaker's lexicon' (p. 268).

Phraseological and formulaic language research evidence shows that the most frequently used words in our languages are only the tip of the expressional iceberg (e.g., Durrant, 2008; Wray, 2013; Wood, 2015; Sinclair, 1991; Martinez, 2011). The extensive use of MWE can also be observed in many spoken examples of language (e.g., good morning, what's up, all right, you know). Most of these everyday phrases can be considered a type of MWE because of the fixed nature of these lexical units and their resistance to any substitution of their component parts.

In Arabic, MWE is a widespread phenomenon. The interests of early Arabic linguists also highlight its unique importance. For instance, in Classical Arabic CA, several scholars paid early attention to MWE and the necessity of studying and collecting these types of formulaic sequences in individual lists or dictionaries. The ancient book on Arabic linguistics *'Arabic aphorisms'* by the early popular linguist

عُبَيد بِنْ شَرِيَّة *'ubayd bin šariyya*, who died in the seventh century, is believed to be the first attempt at data-collection devoted to this phenomenon in Arabic. In modern SA,

MWEs can be observed in most semantic fields and different language genres; several recent corpus-based researchers have provided language data that support the popularity of AMWEs (e.g., Abdou, 2011; Najar et al., 2015). Furthermore, research reveals that the most frequent words in SA usually belong to a more complex network of various AMWEs that dominate the meaning of the core lexemes. For instance, Figure 1.1 shows the underlying complexity of phrases related to the word عين *ʿayn* 'eye').

It is therefore clear that MWE knowledge should not be ignored in any high-quality language processing tasks. The large number of MWEs emphasises their crucial role in the development of most language-related applications.

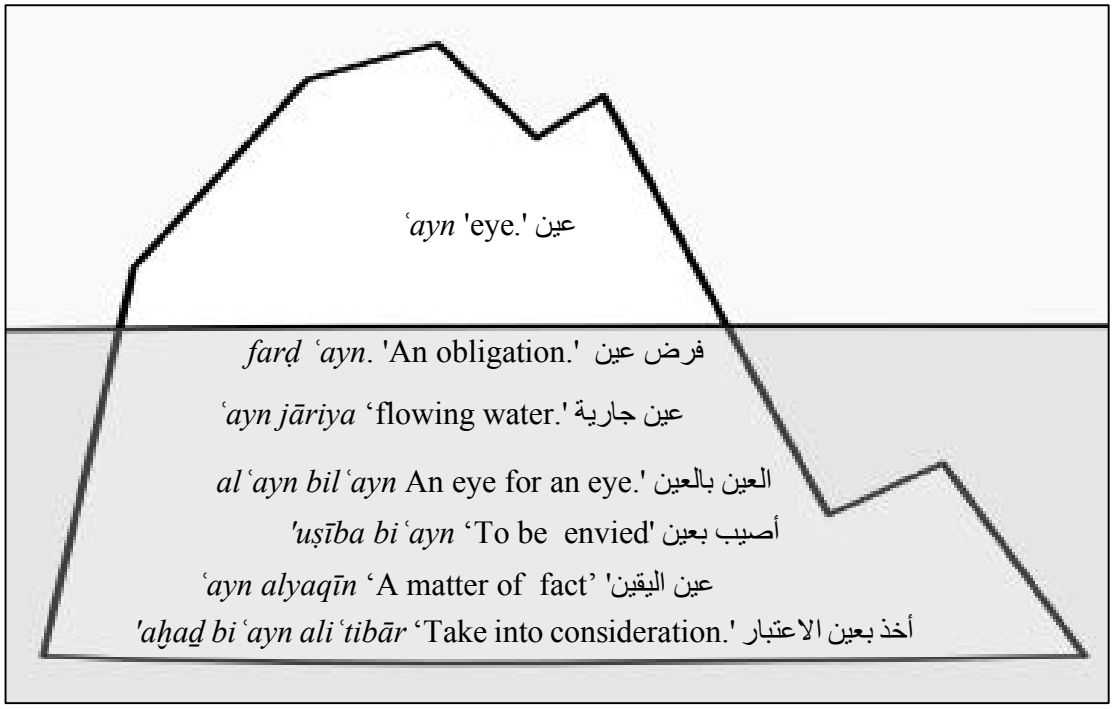


Figure 1.1: Complexity of MWEs related to the Arabic word عين ('eye').

The reason for the ubiquity of MWE and figurative languages in general is illustrated in several studies in the literature. For instance, Dickins et al. (2016, p. 81) introduced the term 'metaphorical force' which explains the capabilities of this type of language and is derived from the strong emotion engendered to satisfy the desire of language users to express their ideas through diverse communicative functions and a range of denotations. Thus, we tend to use and persistently invent many types of MWEs and metaphors to satisfy our emotional needs through various linguistic forms. An

example can be seen in the AMWE (فرض عين *fard 'ayn* ‘obligation'); however, for cultural and emotional reasons, there is strong semantic variation when selecting this expression or merely using the literal alternative single word واجب *wājib* when denoting this specific meaning in SA.

1.2.2 MWEs significance in linguistics and LP

In her comprehensive study on MWEs, Wray (2002) stated that the vital role played by formulaic language means it should be at the centre of any serious study of human language. She emphasised that linguistic knowledge ‘is not only a question of knowing the words that go together into strings but also of knowing the strings of words that go together' (ibid, p. 281). Many phrases used continually in our everyday communications constitute what Sinclair (1991) called “single choices, even though they might appear to be analysable into segments” (p.110). This notion was also stressed by Bollinger (1976), who emphasised that “our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs” (p. 1). The awareness of this phenomenon in human languages therefore began very early, and many researchers have proposed different descriptions and theories for the linguistic behaviour of MWEs and their core role in first and second language acquisition. For instance, Fillmore (1979) correlated language fluency with the ability to control MWEs; he stated that "a very large portion of a person's ability to get along in a language consists in the mastery of formulaic utterances" (p. 92).

Most grammatical theories attempt to partly or entirely accommodate the realm of formulaicity in language systems and consider this phenomenon an essential element of any language structure model. Such theories include Cognitive Grammar (e.g., Langacker, 1991), Construction Grammar (e.g., Brooks and Tomasello, 1999), and Lexical-Functional Grammar (Bresnan et al., 1982). However, an exception can be found in Chomsky’s (1965) universal grammar theory which adopted a generative perspective for explaining grammatical structures and is the theory least tolerant to the idea of associations between lexical items.

Several theorists have attempted to integrate these contrasting theories and have proposed language-processing models that combine an understanding of human language structure systems from two different perspectives. Such attempts can be seen

in the work of Sinclair (1987; 1991), in which he proposes two principles that explain the interactive nature of language use. The first is the open choice principle which, like the Chomskyan account, contends that the creativity of human beings enables them to select individual lexical items and create novel structures based upon abstract universal rules. The second is the idiom principle, which is based on the human selection of different types of sequences that constitute regular strings they have frequently encountered. Sinclair stressed that most linguistic materials could be interpreted in terms of the idiom principle when there is a reasonable justification to do so. Another hybrid model of language processing proposed by Wray and Perkins, (2000) and Wray (2002b), suggests that a dual-system consisting of analytic processing explains the novelty of language use and holistic processing; this is based upon a memorised set of MWEs. However, Sinclair argued that the idiom principle was the superior principle. Wray (2002) also favoured the holistic system of language processing over analytic processing when handling linguistic materials. Although language processing among native speakers can be interpreted simply by either the open choice or analytic processing models, Pawley and Syder (1983) contend that there will still be a large amount of correct grammar that seems to be strange and unlike the authentic native usage of the language. This can be seen in the following quote:

> *''Native speakers do not exercise the creative potential of syntactic rules to anything like their full extent . . . Indeed, if they did so, they would not be accepted as exhibiting nativelike control of the language. The fact is that only a small proportion of the total set of grammatical sentences are nativelike in form – in the sense of being readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd', or 'foreignisms'". (Pawley and Syder, 1983)*

Since Firth's famous (1951) quote, 'You shall know a word by the company it keeps', research in applied linguistics and second language pedagogy has emphasised the major role played by formulaic language and MWEs, particularly in teaching and learning foreign languages. The acquisition of MWEs, beyond the word level in second language learning has been shown to lead to a significant improvement in natural language use and to promote considerable second language fluency. Research

has emphasised the key role of MWE acquisition in the overall improvement of proficiency and fluency in the target language among second language learners (Ellis, 1996; Boers et al., 2006). For instance, after analysing written answers in English given in a foreign language learners' proficiency test, Ohlrogge (2009) found that students with higher grades used MWEs more than those with lower grades. Many MWEs are considered a type of metaphorical language (Dickins et al., 2016) where vivid fluency and proficiency is shown by both native and non- native speakers. Following an intensive analysis of second language (SL) literature, Wray (2002a) found that while in 'the early stages of first and second language acquisition, learners rely heavily on formulaic language to get themselves started', intermediate and advance learners found that 'the formulaic language was the biggest stumbling block to sounding nativelike' (p. 9). Thus, studies in these fields have introduced various theories and language teaching methodologies that stress the critical role of MWEs in second language acquisition. Several criteria for identifying and extracting MWEs have been proposed to ease the process of developing various teaching and learning materials that take this knowledge into account. Other studies have sought to construct different kinds of MWE lists that can be used as a pedagogical tool to facilitate the inclusion of these types of phrases in practical applications.

A large number of researchers have conducted empirical and theoretical studies to gain an in-depth understanding of the linguistic behaviour of MWE from different perspectives. For instance, corpus-driven research findings have demonstrated the essential role played by formulaic language in everyday language (Schmitt, 2010; Ellis et al., 2008; Wray, 2002b; Nesselhauf, 2005). Other studies have found that MWE items play a critical role in conveying various kinds of functions and meanings in language communication (e.g., Biber, Conrad and Cortes, 2004; Hyland, 2008; Dorgeloh and Wanner, 2009; Wulff, Swales and Keller, 2009). For instance, in English and Arabic, several MWEs are used as discourse organisation signposts

من جهة أخرى, *min jiha ʾuḫrā* 'on the other hand'.

Another factor related to the language processing advantages offered by MWEs has been highlighted in several studies that emphasise the easy acquisition of MWE items by native speakers in comparison to standard phrases. In contrast, MWE acquisition is found to be one of the most challenging and difficult tasks for non-native speakers (Siyanova-Chanturia et al., 2011). The complex nature of research in this area has

driven the well-known linguist, Mel'cuk, to describe it as "so difficult, but so appealing!" (Mel'cuk, 1995).

Given the existing theoretical frameworks that attempt to accommodate the phenomenon of MWE in various languages, the current research seeks to present a framework for defining this phenomenon in Standard Arabic and to discover its main linguistic characteristics, laying a theoretical foundation upon which to solve the research problems addressed in this thesis.

1.2.3 MWE significance in computational linguistics and NLP

With the advance of computational tools that enable researchers to explore an unprecedentedly large amount of language data, several studies, mainly in English, have shed light on the significance of the MWE phenomenon and the need to focus on processing these types of phrases by developing various methods for integrating them into language processing tasks. Research conducted by several researchers (e.g., Leech et al., 1983; Smadja, 1993a; Dunning, 1993; Sag et al., 2002) on the development of MWE lists, lexicons, extraction methods, or classification frameworks exemplify the type of early research in this area. Several computational models and lexical resources have consequently been developed for diverse purposes to improve MWE processing tasks. Thus, the vital role played by MWE in computational linguistics and NLP is beyond question; for this reason, a great deal of research has been conducted on MWE from NLP perspectives to improve the computational treatment of this complicated linguistic phenomenon. The inclusion of MWE resources can fundamentally improve the quality of most NLP applications, such as language parsing, information retrieval, machine translation, and foreign language e-learning systems such as Duolingo and Flax projects.[2] Several studies have concluded that accommodating MWE knowledge in NLP tasks is highly beneficial in the reduction of language ambiguity, increases overall precision, and contributes towards naturalising a system's output (e.g., Ramisch, 2015; Carpuat and Diab, 2010b; Rikters and Bojar, 2017).

[2.] For more details visit http://duolingo.com and http://flax.nzdl.org.

Research on MWE in NLP literature can be classified into three areas. The first is MWE computational processing which includes two main subtasks, MWE extraction and MWE identification. The former task aims to find various ways of discovering new MWEs and storing them in lists or lexicons to construct new LRs or enrich existing ones, while the latter aims to automatically identify or annotate existing MWEs in running text to assign them to particular processing tasks. However, in the literature there is a strong overlap between these two subtasks due to the substantial interaction between them. The second research area focuses on creating several types of MWE knowledge bases for use in various NLP or LP applications; such research also encompasses studies on MWE formalisms and computational representations and annotation. The third research area is devoted to embedding MWE knowledge into practical applications to enhance the quality of NLP applications such as machine translation (MT) and language parsing (LP), or to conduct a task-based evaluation of various MWE processing tasks. Although a substantial amount of research has focused on various MWE research problems in the literature, the complexity of these issues and their critical importance in NLP means more research is needed to explore this phenomenon in several languages and from various perspectives. Rayson et al. (2010, p. 3) emphasise that 'despite the considerable effort that has been devoted to the MWE research, there is still a long way to go. The MWE issue is a tough nut, but it needs to be cracked open to further improve NLP and information systems.' This statement remains true over a decade later, particularly for morphologically rich and less-resourced languages. Constant et al. (2017, p. 879) point out that 'An open challenge is how to create lexical resources for under-resourced languages by exploiting comparable data, monolingual resources, or domain specificity'. This thesis will contribute to remedying this gap in knowledge by developing a new AMWE lexicon with computational representations.

1.2.3.1 The need for a computational lexicon of AMWEs

As will be illustrated in detail in section 1.4, this thesis will make additional contributions to MWE research areas in which the aim is to implement and evaluate several AMWE discovery models to create a new MWE LR with a comprehensive formalised system to represent MWE knowledge at various linguistic levels. The availability of machine-readable LRs plays a significant role in improving language processing tasks and this will be illustrated briefly in this section.

Despite the recent and dominant use of statistical methods and artificial intelligence and deep learning techniques in various NLP research tasks, LRs and machine-readable lexicons still play a critical role in the improvement of most NLP tasks. One of the primary applications which demonstrates the need for special processing of MWE is MT, where ignorance of MWE has led to many errors in system output. For instance, in MT between Arabic and English, processing the text without considering MWE knowledge reduces any possibility of producing a high-quality translation output.

Arabic	English Translation
قاصمة الظهر	Back pain

Figure 1.2: Error in google MT output of AMWE ‘قاصمة الظهر qāṣimat aḍḍahr’.

This can be seen in the translation of the Arabic MWE ‘قاصمة الظهر qāṣimat aḍḍahr.' , where tagging this expression as merely a noun/noun sequence and discarding the use of MWE will result in a poor-quality machine translation output, as can be seen in Figure 1.2 which shows the output of a Google MT system. However, this inadequate translation output -which has the opposite meaning to the Arabic expression- could be easily avoided if the system had access to an AMWE knowledge base where the system could map this expression to the closest equivalent single word ‘destroy’, thus leading to better output. This phrase is only one example, there are also many others in Arabic as can be seen in Table 1.1 which shows examples of MT errors caused by inadequate MWE processing. These examples were collected from three sources. The first line presents the En MWE, the second presents the MWE translation by the MT system, and the final line presents the correct translation of MWEs. These types of errors can be easily avoided if the MT system has access to MWE LR.

Table 1.1: Examples of English-Arabic MT errors due to MWE processing.

En source	Waiting to see who had been chosen, we were all **on edge**.
Ar MT	في انتظار لمعرفة من الذي تم اختياره، كنا جميعا **على حافة**.
Ar reference	كنا **نترقب بقلق** من تم اختياره
En source	I could **eat a horse**.
Ar MT	ويمكنني أن **أكل الحصان**.
Ar reference	أنا **جائع جداً**
En source	This mistake was the **final nail in the coffin**.
Ar MT	وكانت هذه المشكلة **الظفر الأخير في نعش**
Ar reference	كان هذا الخطأ هو **الضربة القاضية**
En source	If you suggest a better idea, I am **all ears.**
Ar MT	إذا كنت تقترح فكرة أفضل، أنا **كل الأذنين**.
Ar reference	إذا كان لديك فكرة أفضل، فكلنا **آذان صاغية**
En source	He comes round **once in a blue moon.**
Ar MT	وقال انه يأتي جولة **مرة واحدة في القمر الأزرق**
Ar reference	هو **نادراً** ما يأتي
En source	We are just about **down to the wire** with this project.
Ar MT	نحن فقط نحو **السلك** مع هذا المشروع.
Ar reference	نحن **في اللحظات الأخيرة لإنهاء** المشروع
En source	You should learn to **speak out** in meetings with your boss.
Ar MT	يجب أن تتعلم **التحدث** في اجتماعات مع رئيسك.
Ar reference	يجب أن تتعلم أن **تطرح رأيك بجرأة** أمام مديرك
En source	The company investment funds to **Land Windfall**.
Ar MT	الشركة صناديق الاستثمار إلى **لاند وينفال**.
Ar reference	حاز صندوق الاستثمار في الشركة أرباحاً **هائلة**

Another primary benefit of creating LR is the opportunity these lexical resources provide to explore and examine the behaviour of several linguistic phenomena. This will provide sufficient data to answer the long-standing question as to how our language functions in its various manifestations. Statistical methods offer little in this area in comparison to the contributions of LRs. It is widely known that language data is distinct from many other sorts of data. Therefore, the transfer of several statistical concepts and applications from other research areas should be conducted with caution and should consider the core characteristics of linguistic data. For instance, in his famous paper, Kilgarriff (2005) contends that 'when we look at linguistic phenomena in corpora, the null hypothesis will never be true. Moreover, where there is enough data, we shall (almost) always be able to establish that it is not true'. He also states that a better result would be obtained if more time was spent on enriching existing

LRs with rich annotation rather than conducting repetitive statistical experiments. Developing MWE LR with rich annotation plays a significant role in the improvement of several MWE computational tasks as these lexicons can be used to enhance MWE discovery and identification models (Constant et al., 2013; Bejček et al., 2013).

Although creating linguistic LRs is a costly, labour intensive, and time-consuming construction process, many statistical methods still base their results on reference corpora which have to be constructed and annotated in the same way as creating linguistic LRs. The aim is not to prove that one method is better than the other but to show the significance of developing linguistic LRs to improve most NLP tasks. As can be seen in several current NLP studies, a hybrid model is adopted that takes advantage of both linguistic and statistical methods.

The final point in this section is related to the importance of AMWE research. Several researchers in the NLP Arabic research community have highlighted the imperative for developing different kinds of AMWE LRs for use in NLP applications. For example, Bar, Diab and Hawwari (2014) pointed out the lack of comprehensive Arabic MWE resources, particularly those that can be integrated easily into practical applications. Ebd-alrzaq (2007) states that most Arabic NLP tools are still based on listings of single orthographic words due to the absence of well-developed AMWE resources. Although the importance of English MWEs has been acknowledged by many researchers in the field of NLP, as evidenced by a large number of studies and dedicated conferences and workshops, the theory of Arabic MWEs is still underdeveloped. In comparison with English research, Arabic computational lexicography is still in the embryonic stage, and there is an urgent need to enhance Arabic lexicographic research through advances in several computational methods in NLP. Another research study by Abdou (2011, p. 233) proposes 'developing an electronic database of Arabic idioms that includes information on their linguistic behaviour, particularly their variation potential … Indeed, (corpus-based) investigations of Arabic idioms and Arabic phraseology in general that are synchronic or diachronic in nature are much needed for both theoretical and practical purposes.'

In summary, all the research discussed illustrates the critical need to study AMWE from both theoretical and practical perspectives, and that is what the current research project therefore aims to do. The importance of this research lies in a set of factors related to the vital importance of integrating MWE into NLP and other linguistic

applications. Lack of knowledge as to how to handle MWEs in any language-related tasks will hamper the processing of many languages which will undoubtedly have a negative impact on their final output quality.

1.3 Task definition

A survey of MWE definitions and terminology along with the conceptual framework adopted for AMWE is presented in chapter 3. Hence, this section focuses only on highlighting several vital issues related to the context and scope of the thesis and focuses on describing the main objectives, research questions, and the contributions that will be made.

1.3.1 Research context and scope

Building a comprehensive MWE LR is a long-running task that is likely to need a dedicated multidisciplinary work team with adequate funding and other related resources. Therefore, it is essential to concede that the current project is the result of one individual's work within a set time limit. This clarification is essential in explaining the boundaries of the project. Thus, in the thesis the intention is not to create an exhaustive AMWE LR but to focus on achieving specified research objectives. The term 'Arabic' refers to one variety of the language called Standard Arabic (SA), which will be described in section 3.2 of this thesis. Furthermore, the use of any commercial LRs and tools to which the researcher does not have access will be excluded. The following subsections describe the primary objectives and questions of the thesis.

1.3.2 Research objectives and questions

The following are the core objectives of the research:

To propose a theoretical framework for describing AMWE criteria and concepts, and highlighting their distinctive linguistic properties at various levels of analysis.

To develop a computational corpus-informed AMWE lexicon that can be incorporated into various Arabic NLP applications.

To construct a model for describing and encoding AMWE lexical entries at different linguistic levels (morphological, syntactic, lexical, and semantic).

To determine the information and annotation that will best serve the needs of language-related and NLP applications.

To implement an overall model for AMWE extraction that will best suit the primary objectives of this research.

To explore the feasibility of creating an extensive AMWE LR by conducting several AMWE extraction experiments and constructing a large lexicon consisting of various types of AMWE entries with rich linguistic annotations.

1.3.3 Research questions

Based on these objectives, the following are the central research questions that will be addressed in this thesis.

RQ1: What types and definitions of AMWEs should be given priority in light of the research problems addressed in this study?

RQ2: How can lexical units of the type defined in RQ1 be discovered using computational extraction models?

RQ3: What are the standards and best practices for linguistic annotations and computational representations of AMWE knowledge at various linguistic levels?

These questions summarise the core problems that will be addressed in this project, and include several detailed sub-questions as follows:

What are the core criteria for defining the targeted AMWEs?

What are the linguistic characteristics that distinguish AMWE from other lexical units and various types of language sequences?

What is the best overall architecture for discovering these types of AMWE from the corpora?

What are the most relevant information and linguistic annotations that should be included in the targeted computational lexicon of AMWEs?

1.4 Thesis contributions

The novel contributions made by this thesis can be classified into three types of AMWE computational processing tasks: AMWE extraction, evaluation, and building new large AMWE LRs with an in-depth formalised model representing AMWE

knowledge at various linguistic levels. These will be described in the following subsections.

1.4.1 A theoretical framework for AMWE

The first task is to present a detailed framework for describing AMWEs and illustrating their various linguistic properties and varying potentials. This is an essential step in solving the research problems stated in this thesis. The linguistic description of AMWEs will provide a beneficial contribution that can be utilised by various related studies in AMWE research.

1.4.2 AMWE discovery models

One of the primary objectives of the current study is to develop an innovative hybrid model and framework for the discovery of AMWEs from various types of large SA corpora. Moreover, the research aims to implement several evaluation methods that will validate the proposed extraction approach and measure its efficiency and usefulness.

1.4.3 Language resources

The AMWEs lexicon, which is the ultimate aim of this project, will be of use to interested researchers, Arabic teachers, and learners. This lexicon also can be integrated into several NLP applications to eliminate language ambiguities.

1.4.4 Representations and a formalising framework for describing AMWEs

The current project aims to construct an intensive framework that formalises AMWE knowledge at different linguistic levels (e.g., morphology, syntax and semantics) to facilitate the integration, usability, and scalability of the developed AMWE LR. This will have a positive impact on the process of embedding MWE knowledge into practical applications.

1.5 Thesis organisation and published work

1.5.1 Organisation

This thesis consists of eight chapters. The first presents an introduction to the project and describes the motivation underlying this research as well as stating the research questions and objectives. Chapter two provides a survey of relevant works under three core research areas: MWEs extraction methods, MWE LRs, and computational representations and formalisms of MWE knowledge. Chapter three presents a general background to MWEs and their linguistic characterisations with a focus on producing a detailed framework for AMWEs. Chapters two and three address RQ1, which will provide the foundation for the next research study reported in this thesis. The research experiments reported in chapters four, five and six address RQ2. Chapter four presents an experiment related to the development of gold standard reference lists of AMWEs that can be used later as evaluation datasets. Chapter five and six present a series of experiments related to the implementation of multiple AMWE discovery models used for extracting and evaluating various types of AMWEs. Chapter seven addresses RQ3 by providing a comprehensive and formal model for representing various types of AMWEs. Finally, chapter eight concludes with a summary of the research findings, challenges, and potential future work.

1.5.2 Published work

Within the time constraints of this thesis, and with the help and encouragement of my supervisor Prof. Atwell, parts of the work presented in this thesis have been published as follows:

Chapter 4:

Alghamdi, A. 2015. The development of an Arabic corpus-informed list of formulaic sequences for language pedagogy. In: The eighth international Corpus Linguistics conference., University of Lancaster, UK.

Alghamdi, A. and Atwell, E. forthcoming. Constructing a corpus-informed Listing of Arabic formulaic sequences for language pedagogy and technology. Accepted paper submitted to the International Journal of Corpus Linguistics.

Alghamdi, A. and Atwell, E. 2018b. An Arabic corpus-informed list of MWEs for language pedagogy. In: O. L. Dong, J. Lin, W. Xiao, M. Geraldine and P.-P. Pascual,

eds. TALC 2018 13th Teaching and Language Corpora Conference. Cambridge, pp. 38–41.

Chapter 5:

Alghamdi, A. and Atwell, E. 2016. An empirical study of Arabic formulaic sequence extraction methods. In: LREC'2016 10th Language Resources and Evaluation Conference. Portorož, Slovenia.

Alghamdi, A. and Atwell, E. 2016b. Towards a Computational Lexicon for Arabic Formulaic Sequences. In: The International Conference on Information and Communication Technologies. IRCAM institute, Rabat, Morocco.

Chapter 7:

Alghamdi, A. and Atwell, E. 2017b. Towards Comprehensive Computational Representations of Arabic Multiword Expressions. In: R. Mitkov, ed. Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, November 13-14, 2017, Proceedings [Online]. London: Springer International Publishing, pp. 415–431.

1.6 Summary

In this chapter, a general introduction to MWE research was presented and several vital issues related to the context and scope of this thesis were highlighted. This was followed by a description of the core reasons that constitute the primary motivation for conducting this research. The main tasks of the thesis were then outlined and the primary research objectives and questions specified. The thesis structure was then described and outlined along with references to the work already published from this project.

2 Literature Review

2.1 Introduction

The literature review in this chapter has been organised according to the main research questions and objectives of the thesis. The aim is to develop computational models for extracting multiple types of AMWEs to create a lexicon with detailed representation and formalism which covers various levels of linguistic description. This review is therefore divided into three main sections. The focus in sections 2.2 and 2.3 will be on reviewing the discovery, extraction, and evaluation of MWE knowledge in the literature. Section 2.4 will then discuss related existing MWE LRs and computational lexical representations with a particular focus on AMWE studies when available.

First, however, it is important to provide an overview of the essential research areas within the realm of MWE in NLP. The idiosyncratic nature and overlapping boundaries of these types of expression have impelled researchers to investigate this phenomenon from various perspectives which include but are not limited to lexicology, language pedagogy, morphology, syntax, and semantics. Nevertheless, as mentioned previously in section 1.2.3, most NLP research on MWE can be classified into one of these main areas, which comprise the following sub-classifications:

Research on the computational processing of MWE which primarily includes discovery and identification tasks.

Evaluation studies that suggest and implement multiple evaluation methods for MWE processing tasks.

Developing an MWE lexicon and other LRs for various applications.

The representation of MWE knowledge based on multiple lexical and formalism models.

Application-oriented research which focuses on integrating MWE knowledge into various NLP tasks such as developing MWE- aware LP or MT systems.

2.2 MWE Discovery methods

This section presents a brief survey of research on MWE extraction and discovery methods. Several classifications can be used to organise research in this section based on the adopted view of typology methods suggested in the literature. For instance, MWE extraction methods can be classified according to a historical timeline of research development or they can be classified based on the type of performance in the models, such as manual, automatic, or supervised and unsupervised discovery models.

However, in this review, the classifications of extraction methods based on the primary approaches used in most NLP tasks will be adopted, which are statistical, linguistic, and hybrid approaches. Thus, this review is divided into three sections, based upon the main approaches to the extraction of MWE. It is worth noting that there are no strict classifications of MWE discovery methods in the literature. This is because there is usually no clear-cut distinction between extraction methods in real applications. A great deal of overlap is therefore anticipated at various levels of processing given the dominant use of the specific MWE discovery approach.

It is first important to illustrate what is meant by the MWE extraction or discovery model in the context of this thesis. The MWE discovery model primarily denotes the process by which text corpora are selected and then an AMWE extraction model applied to the textual data to discover multiple types of AMWE in various morphosyntactic patterns and semantic domains. Thus, the final output of this process is a list of many lexical sequences that can be later evaluated or filtered by experts to create or enhance MWE LRs.

Research in this area dates back to the 1960s, since when several papers have been published on MWE and various methods for discovering their multiple patterns from corpora (e.g. Stevens and Giuliano, 1965; Berry-Rogghe, 1973; Atwell, 1988; Choueka, 1988; Leech et al., 2001; Leech et al., 1983; McEnery et al., 1997). Most of the early research in this area focused primarily on experimenting with different computational methods for extracting MWEs or on conducting a comparative evaluation of knowledge-based and statistical extraction models, primarily on English and other European languages.

An example of this early research can be seen in the work of Leech et al. (1983) through their work on the development of a LOB[3] corpus tagging project. Multi-word or ditto tags were first created for ‘a sequence of two or more orthographically separate "words" functioning as a signal lexical item (e.g., ‘no one’, ‘so that’). This method is very beneficial in the automatic extraction of immutable phrases from a POS tagged corpus, but this is not the case when the goal is to discover multiple flexible constructs of MWEs, especially in morphologically rich languages which have more complex morphosyntactic systems and possible variants of MWE.

These early attempts at using computational methods to discover linguistic patterns continued and various techniques and models have since been suggested in the literature (e.g., Dias et al., 2000; Bartsch, 2004; Krenn, 2000; Todiraşcu et al., 2008; Piao et al., 2003; Sag et al., 2002, among others). Most studies have mainly been applied to English due to early access to machine-readable LRs and the interest of a large research community in corpus linguistics and NLP. The complexity of MWE extraction tasks means this issue still poses various open research problems; further research is therefore required to remedy knowledge gaps in this area. Piao et al. (2003) point out that, despite a substantial amount MWE extraction research, ‘efficient extraction of MWEs still remains an unsolved issue'. This largely remains the case although there have been remarkable developments in this research area for English and other European languages. However, there is still a need for further experiments and research, particularly for morphologically rich languages such as SA. This need is also supported by the fact that MWE is a linguistic phenomenon that continually changes and many new types and structures of MWEs emerge on a regular basis. It is important to note that in MWE extraction research there is a circular relationship between the MWE definition adopted and the extraction methods implemented. Hence, every theoretical framework for MWEs leads to the selection of a specific approach in MWE extraction tasks.

[3] This is an abbreviation for (Lancaster-Oslo/Bergen Corpus of British English).

2.2.1 Knowledge-based approach

In the literature, this approach is also termed a symbolic, linguistic, and phraseological approach to MWE extraction. Research following this approach emphasises the crucial role of linguistic processing components and characteristics of MWEs in the extraction model. The definition of MWE, according to this methodology, is based on the structural relations between the lexical items in MWE. The works of several researchers (e.g., Bartsch, 2004; Cowie, 1998; Mel'ćuk, 1998) represent an understanding of MWEs from the linguistic perspective. For instance, Bartsch (2004) defines collocations as, 'Lexically and-or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other' (p. 76). This definition illustrates the core role of structural relations in identifying collocations between the lexical items. An alternative definition, embedded within the meaning-text theory proposed by Mel'cuk, is considered one of the most popular definitions of collocation and MWE within the linguistic approach. The following paragraph explains the concept of collocations based on this theory.

> *(16) Let AB be a bipartite language expression, where A and B are lexical items of the language L, and let 'S' be the meaning of AB, 'A' the meaning of A, and 'B' the meaning of B. The expression AB is a collocation if the following three conditions hold:*
>
> *(i) 'S' ⊃ 'A' (the meaning of S contains the meaning of A);*
>
> *(ii) A is selected by the speaker in a regular and non-restricted way;*
>
> *(iii) B is not selected in a regular and non-restricted way, but depends on A and the meaning of 'S' to be expressed (Mel'cuk, 2003).*

This concept emphasises the impact that the relationship between the collocation items has on its meaning. According to this theory, the base word in the collocation plays a significant role in determining its meaning. The lexical function language-modelling tool, based on Meaning-Text theory, has had a substantial impact on NLP research and has been applied to various NLP applications such as MT and language parsing. (e.g. Dorgeloh and Wanner, 2009; Pal, Naskar and Bandyopadhyay, 2013). The following subsections briefly outline the core linguistic components that can be embedded in MWE discovery models at various processing stages.

2.2.1.1 Tokenisation

Following the essential normalisation tasks[4] used in most NLP tasks, tokenisation is a vital step in any AMWE extraction model because it eliminates noisy data and is also a prerequisite for other basic linguistic tasks such as lemmatisation and POS tagging. These ultimately assist in the improvement of several statistical functions such as the frequency counts of the text. The primary objective in this task is to split the textual strings into several clusters which represent various morphemes and affixes based on a specific tokenisation scheme; the output thus consists of multiple types of token that represent different morphological units. As will be described in section 3.2.1, SA has several distinctive properties that emphasise the significance of this non-trivial task in the AMWE extraction model.

For instance, the right tokenisation of multiple affixes in SA enables the recognition of many AMWE that are not space delimited words but instead consist of one textual string. as can be seen in the example[5] below:

و | بـ | الـ | تالي

wa.bi.ttālī [6]

Therefore

Splitting the text into parts at the sentence level of analysis can be considered another type of tokenisation, also called text chunking or shallow syntactic analysis, and is supported by several NLP toolkits. In SA, many tokenisation schemes can be found which start from a simple scheme based on the use of white space or punctuation as separator marks by implementing regular expression functions and progress, to other, more complex, tokenisation systems which involve several morphological disambiguation tasks that enable the tokenisation tool to split the text based on intensive and complex morphological models.

The selection of an appropriate tokenisation scheme is usually based on the requirements of each NLP task. Although this task received early attention in the

[4] The common SA normalisation tasks are presented in section 3.2.1.1 of this thesis.

[5] More instances of one-string AMWE are presented in section 3.4.4.5.

ANLP research community and an enormous amount of research has been devoted to developing different methods for improving its accuracy, it is still considered an open research problem. This is particularly the case for morphologically rich languages which still require more advanced models to eliminate multiple types of tokenisation error.

2.2.1.2 Lemmatisation

Lemmatisation is another core linguistic component that enhances and improve the AMWE extraction process: the use of a lemma strongly affects the statistical analysis and frequency information extracted from the corpus. The count of all inflectional forms instead of the core lemmas of MWE candidates leads to redundant and inaccurate statistical data about various linguistic units. This task is based on the output of a previous tokenisation task which enables the tool to identify all inflectional or derivational forms which can then be mapped to their root or core lexeme.

This is a significant step, particularly for morphologically rich languages which have many related inflectional forms for each lexeme. Statistical MWE extraction research has found that using the cumulative frequency of a specific lemma and all its inflected forms has a significant advantage over merely counting the frequency of each inflected form (Evert and Kermes, 2002; Evert et al., 2004). However, the lemmatisation task in most available ANLP toolkits is far from established due to the complex morphological system of SA. Thus, the adoption of lemmatisation in an AMWE extraction model should be applied carefully to avoid any unwanted or misleading outputs.

2.2.1.3 Diacritisation or vocalisation

This process refers to the process of adding short vowels, nunation, and gemination or syllabification marks to SA text to improve the morphological analysis. This is because different diacritisations of words usually leads to various morphological and lemmatisation results[7]. Dediacritisation, which involves removing these marks, is another pre-processing task utilised when the aim is to normalise the text or reduce

[7] Several examples are provided that show the effect of various diacritisations on the linguistic analysis of SA text in section 3.2.1.1.

the complexity of morphological analysis. This task plays a vital role in reducing morphosyntactic disambiguation tasks in SA (Habash, 2010a). Much research has been conducted on automatic and semi-automatic diacritisation tasks. Examples can be seen in several research studies (e.g., Shahrour et al., 2015; Abandah et al., 2015; Obeid et al., 2016; Azmi and Almajed, 2015) which have mostly yielded high-precision results. Most of the work in the ANLP research area has focused on the simplified diacritisation task which avoids the processing of the word final diacritics because, in most cases, they are used to indicate the syntactic case of the words based on the morphosyntactic context. This advanced linguistic analysis requires in-depth syntactic parsing which is still a challenging problem in ANLP research.

2.2.1.4 Part of speech tagging

Adding a POS tag to each token is considered an essential phase in linguistic processing. However, this is a long-standing field of research in ANLP which faces both enormous challenges and opportunities. The primary source of complexity of this task in SA is the extremely wide variation in the number of POS tagsets, which ranges from three possible core tags to theoretically more than 330,000 potential tags based on various morphosyntactic features (Habash and Rambow, 2005a; Habash, 2010a). Thus, the comparative evaluation of POS taggers in SA is a challenging task. Nevertheless, most computational toolkits available for SA depend on a reduced POS tagset in their morphological analysis which assists considerably in achieving adequate accuracy of output (e.g., Attia, 2006b; Saad and Ashour, 2010; Buckwalter, 2002; Pasha et al., 2014; Sawalha, 2011)[8].

2.2.1.5 Parsing

Syntactic analysis is another linguistic process that refers to modelling the syntactic relation system between various tokens in the textual data and retaining all the morphosyntactic information of the sentence to produce a detailed syntactic analysis based on multiple linguistic frameworks and syntactic theories. In SA, this task overlaps substantially with the morphological analysis because several syntactic

[8] More details about a specific ANLP toolkit such as MA and SAP will be presented in the AMWE extraction experiment.

relations are indicated by internal modifications in the cliticization morphology[9]. Although several Arabic syntactic treebanks can be found (e.g., Dukes and Buckwalter, 2010; Maamouri and Bies, 2004b; Hajic et al., 2004; Habash et al., 2009; Dukes et al., 2010)), SA still lacks an open source deep morphosyntactic parser which takes input text and generates comprehensive morphosyntactic parse trees with adequate levels of precision.

Alternatively, shallow syntactic parsing, which is related to text chunking tasks based on POS tagging or specific orthographical marks such as punctuation, can be used and is supported by most ANLP disambiguation toolkits (e.g., Pasha et al., 2014; Manning et al., 2014). Figure 2.1 presents an example of a complex rich parse tree from the Prague Arabic dependency treebank.

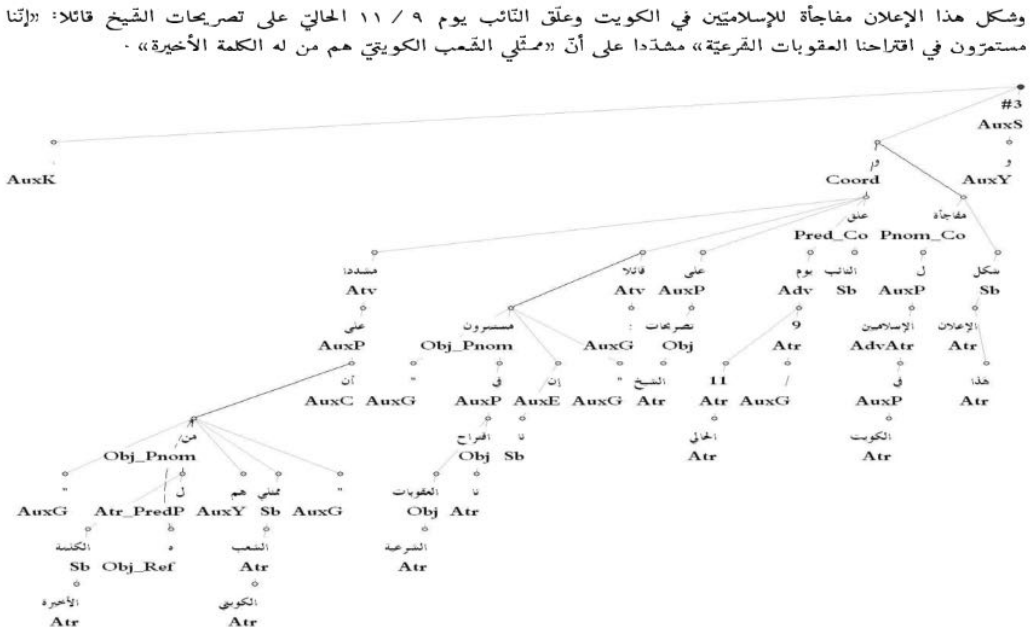


Figure 2.1: Example of SA syntactic parse tree (Hajic et al., 2004, p. 5).

In MWE extraction, deep syntactic parsing enables the discovery model to learn various constructions by retaining the information related to the syntactic modifications and relations between diverse POS combinations. Research on MWE has found that shallow and deep parsing has a positive impact on the final extraction

[9] In section 3.2.1.4, a brief description of the SA morphosyntactic structure will be presented with examples.

outputs and increases the coverage of the extraction model (e.g., Seretan, 2011; Pecina, 2010). Other research has implemented the chunk-based or shallow syntactic approach in the process of collocation identification which involves detecting various ranges of syntactic structures that include PP_Verb (Begoña Villada Moirón, 2004; Krenn and Evert, 2001), Verb_Noun (Wu and Zhou, 2003; McCarthy et al., 2003),Noun_Noun (Bergsma and Wang, 2007), and Adjective_ Noun (Seretan et al., 2004).

Seretan (2011) presents a comprehensive framework for syntax-based collocation extraction based on deep syntactic parsing and provides an example of research following this methodology in MWE acquisition. Using the Fips Multilingual Parser, Seretan developed an extraction systems architecture that consists of two main phases; candidate identification, based on the syntactic structures, and candidate ranking, based on syntactic parsing findings and the use of association measures (AMs). Figure 2.2 shows an example of a parse tree generated by Fips (Seretan, 2011, p. 64).

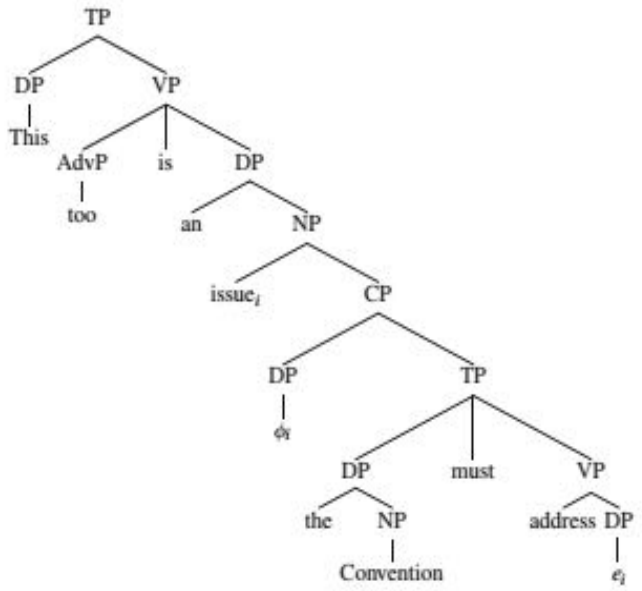


Figure 2.2: Example parse tree for the sentence 'This too is an issue the Convention must address' (Seretan, 2011, p. 64).

The syntactic parsing in this research was based on the concepts of lexical functional grammar (LFG) which determine the system adopted for relations between various syntactic constituents. In this research the term 'collocation' was used, and the utilised concept of collocation encompasses all 'lexical combinations that are: (a)

prefabricated, (b) arbitrary, (c) unpredictable, (d) recurrent, and (e) unrestricted in length' (ibid, p.27).

Although in this study both empirical and frequency data was used, the research mostly relied on the linguistic analysis when selecting potential MWE candidates, as Seretan (2011 p. 66) explains that 'the main criterion for selecting a pair as a candidate is the presence of a syntactic link between the two items'. In the extraction process, several constraints were applied such as excluding proper nouns and auxiliary and modal verbs. Hence the lexeme candidates had to be common nouns or ordinary verbs. Table 2.1 shows examples of extracted candidates along with their POS combinations and syntactic relations.

Table 2.1: Examples of extracted collocations items (Seretan, 2011, p. 67).

Collocation	POS combination	Syntactic relation
Wide range	Adjective-noun	Head-modifier
Work concerned	Noun-adjective	Head-modifier
Food chain	Noun-noun	Head-modifier
Fight against terrorism	Noun-preposition-noun	Head-modifier
Rule applies	Noun-verb	Subject-verb
Strike balance	Verb-noun	Verb-object
Point out	Verb-preposition	Verb-particle

These examples show that there were no constraints in the morphosyntactic patterns. This reflects what is intended in the current research where multiple types of morphosyntactic combinations will be included.

Attia (2008) applied LFG theory to tackle the problem of morphosyntactic ambiguity in SA by building a language parser using the Xerox Linguistics Environment (XLE),[10] which was developed as a platform for writing rule-based grammar systems for various languages within the LFG framework. However, part of Attia's research focused on handling specific types of AMWE to reduce language ambiguity in the output of the morphological transducer[11]. To achieve this, Attia built a specialised

[10] XLE is a tool for parsing and generating Lexical Functional Grammars. For more details, see: http://ling.uni-konstanz.de/pages/xle.

[11.] The term 'transducer' means 'A kind of automaton consisting of a finite number of states connected by transitions. Some states are initial, some final, and the transitions are decorated by symbols. An automaton accepts a string of symbols whenever one can begin at an initial state and follow transitions

two-sided MWE transducer that involved fixed and semi-fixed expressions by using a finite state regular expression. Due to the constraints of transducers, all nouns that allow external elements to intervene and all verbal MWEs were excluded. AMWE were collected using a manual and semi-automatic corpus concordance tool. All fixed compound nouns were encoded in a list of finite state regular expressions, as can be seen in the following example:

- ["+noun" "+masc" "+def"]: .{الأمن}sp {حفظ} (ibid, p79).

Semi-fixed MWEs which might undergo morphological or lexical variations were added to the list with specific tags that demonstrated their different variations, as can be seen in the following example which can be used with or without the determiner 'ال':

- ["+noun" "+masc"]: .{سلاح}("+def":{ ال}) {ال}{نزع}

In Attia's study, several LFG rules were written to cover different types of AMWEs and their potential variations. He found that the integration of MWE knowledge during the processing and pre-processing phases in the morphosyntactic analysis resulted in a considerable reduction in the ambiguity of the parsers' output Attia (2008 p. 88) concludes that 'when MWEs are properly dealt with, they reduce parse ambiguities and give a noticeable degree of certitude to the analysis'.

2.2.1.6 Morphosyntactic patterns

This linguistic technique used in several MWE discovery models is based on using regular expressions to extract multiple types of selection morphosyntactic patterns which represent various templates and POS combinations from linguistically annotated corpora. The generated output of this process is a list of patterns and their surface forms which can be used later in extracting multiple MWE instances.

The work of several researchers (e.g., Justeson and Katz, 1995; Hearst, 1992; Hearst and Hearst, 1998) exemplify the type of research in which this method has been utilised, particularly in terminology extraction research. Another study on Italian MWE by Castagnoli et al. (2014) implemented this method by using a predetermined

designated in the string, arriving at a final state with no further elements to process. Generation is similar' (Bussmann, 2006, p. 411).

list of POS patterns from 19 bigrams and trigrams[12] to extract MWEs which contain at least one adjective. Table 2.2 provides examples of the POS patterns used for extracting MWEs.

Table 2.2: Examples of POS-patterns used for MWE discovery (Castagnoli et al., 2014, p. 58).

POS patterns	Examples	Translation
ADJ ADJ	stanco morto	dead tired
ADJ CON ADJ	vivo e vegeto	live and kicking
ADJ NOUN	prima classe	first class
NOUN ADJ ADJ	prodotto interno lordo	gross national product
ADJ CON ADJ	pura e semplice	pure and simple
VER ADJ	uscire pazzo	to go crazy

The study was based on a newswire contemporary Italian 300M corpus which was annotated with a POS tagger. The extracted lists of MWE were then classified according to the position of the adjective in the sequences: initial, middle, and final. Based on their findings, Castagnoli et al. concluded that the predetermined list of POS patterns is an effective method for exploring MWE knowledge, especially if the selection patterns involve a wide range of common MWE constructs. In the AMWE extraction experiments conducted in the current research, most of these linguistic components and multiple sources will be used to select the most predictive morphosyntactic patterns of MWEs in SA.

2.2.1.7 Gazetteers

Gazetteers is another linguistic method that is based on the use of existing MWE repositories to find similar sequences in the text. In this method, platform MWE items must be encoded in specific ways according to environmental standards and the rules of systems. However, most research using this method has focused on MWE identification tasks,[13] particularly MWE studies on fixed expressions and named entity recognition. The main limitations of this method lie in the inadequate handling of flexible and discontinuous types of MWE. Several researchers have attempted to

[12] In this study the researcher used their intuition and lexicographic sources to select POS patterns.

[13] In section 2.3 of this chapter, a distinction was made between MWE extraction and identification tasks. However, in the context of this thesis, the concern is with the former.

create methods for facilitating the list development process, as can be seen in Maynard et al.'s (2004) research on creating an automatic tool for collecting gazetteer lists for use in the GATE[14] NLP platform (Cunningham, 2002).

2.2.1.8 Translation

Machine and manual translation[15] used in MWE LR building assumes that common MWEs found in one language might have corresponding MWEs in other languages, thus MWE LRs can be translated into other languages. One of the main advantages of this method is that it eliminates the MWE LR creation process by quickly building a translated copy of existing MWE lexicons.

Several researchers have attempted to translate the English version of word-net to create respective LRs for other languages, such as Slovene (Vintar et al., 2008) and Arabic (Attia et al., 2010). Other researchers have used the translation method for creating semantic LRs (Piao et al., 2017; El-haj et al., 2017) to assist in the development of semantic taggers similar to the original English semantic analysis system (USAS) (Rayson et al., 2004).

The translation of MWEs is considered a highly beneficial way of creating and extending specific types of MWE such as named entities and somewhat compositional MWEs. However, the main drawbacks of this method are a lack of high-quality updated bilingual lexicons for most languages, which negatively affects the translation output. Furthermore, a reliance on this method leads to the loss of precious information about many types of MWE knowledge specifically related to expressions that do not have corresponding MWEs in other languages. This is a serious point to consider given that most MWE knowledge is regarded as a language-dependent phenomenon in that it is intensively related to the context and culture of the targeted language. Thus, in most cases, non-compositional MWEs lose their original meaning when they are translated into other languages.

---

[14] GATE is an open source general architecture software capable of solving NLP problems. It was developed at Sheffield University, UK, which can be downloaded from https://gate.ac.uk

[15] Translation is used directly to create MWE LRs; however, in section 3.2.1.9 translation is used as a type of semantic method to discover opaque MWE items.

2.2.1.9 Semantic or Non-compositionality detection methods

Several semantic techniques have been used for extracting MWEs. For instance, two semantic methods that have been widely used for recognising MWEs with a high degree of non-compositionality and fixedness are non-substitutability and non-literal translatability. The former has been used for extracting fixed MWEs that are resistant to any types of variability, which means their components cannot be replaced with any other synonyms or alternative lexical items. The latter method has been used to identify phrases that cannot be literally translated into other languages, which means a word for word translation method cannot be used to render them correctly. However, these two semantic features of MWEs are used in many studies as part of the linguistic elements involved in extracting or classifying MWEs.

Other research on semantic MWE extraction implements semantic field taggers to extract MWEs. For instance, Piao et al. (2003) used the USAS system to implement an experiment on MWE extraction from a domain-specific newspaper corpus related to court events. The semantic tagger assigns semantic field tags to MWE candidates based on the most relevant meaning of the extracted MWEs. The retrieved list consists of 4,195 MWEs which were then subjected to manual checking and reduced to 3,792 items. The extraction precision was 90.39%, and the final MWE list was classified into several categories based on semantic tags. The "names and grammatical words" class was the dominant semantic category with 1635 MWEs followed by the "time" category with 459 items. The MWE length includes expressions from 2 to 6 words although the majority of extracted MWEs were bigram constructions.

2.2.1.10 Other linguistic techniques

Several other linguistic methods can also be found in the literature that are either used purely or, as in most cases, with other statistical methods. For instance, Bourigault (1992) employed several linguistic means for extracting MWE terms by developing the LEXTER system which targets the retrieval of different types of MWTs. The system consists of two main phases. First, the text is analysed to identify the phrase borders by comparing various constructions to predetermined grammatical patterns: a special tool is then used to exclude all structures that do not match any predetermined patterns. The second stage involves parsing the maximal-length noun phrases. In this phase, the system analyses potential MWE candidates based on a

rules-based parsing module which ultimately leads to the extraction of MWE constructions that are most likely to be considered terminological units.

Another study by Heid (1998, p. 12) used several linguistic components in building a German MWE extracting tool that consisted of 4 main stages, which were as follows:

Find single-word term candidates and relevant morphemes in single-word term. (including compounds).

Find all compounds with relevant morphemes

Find multiword terms.

Apply filters for “term status".

A comprehensive list of possible MWE linguistic extraction methods is beyond the scope of this review. Thus, only brief insights will be provided into the common and related linguistic components frequently used in various types of MWE discovery models.

2.2.2 Data-driven approach

In the literature, this approach is also described as the statistical, distributional, or frequency-based approach to MWE extraction. It principally concentrates on modelling the statistical behaviour of MWEs in various language contexts. Extraction methods based on this approach were among the earliest techniques used, especially in NLP literature (e.g., Stevens and Giuliano, 1965; Berry-Rogghe, 1973; Smadja, 1993; Sag et al., 2002).

In the MWE discovery process, these methods primarily focus on using frequency counts and probabilistic distribution context information for words or tokens in the text to determine statistically notable sequences based on various statistical criteria. Firth's definition of collocation was one of the earliest descriptions of collocations or MWE to emphasise the statistical features of these types of lexical units; he defined a collocation of a given word as: …statements of the habitual and customary places of that word’ (Firth, 1957, p. 181).

Another definition, which also adopts a statistical view of MWE, was given Sinclair (1991 p. 170), who defined collocation as; ‘the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a

maximum of four words intervening'. This concept of MWE is the most dominant and influential, particularly in NLP and pedagogical MWE literature, because statistically based methods are usually considered straightforward language models to apply in practice. They also take advantage of the computational power available for the statistical processing and analysis of big data.

However, depending purely on statistical methods has several drawbacks. For instance, Moirón (2005a) states that the amount of noise in statistically extracted MWE lists is often considerably higher than in other methods. Furthermore, the AMs used in this approach work primarily with bigram MWE candidates which might limit the extraction of more extended sequences. However, studies in this area have implemented several tactics to overcome these limitations, as will be described briefly in the following subsections, which review two statistical techniques utilised frequently in MWE extraction, the n-gram and AM models.

2.2.2.1 N-gram model

The n-gram is a probabilistic language model that has been used in MWE extraction experiments in different settings and various language domains. It was initially based on the work of Shannon (1916-2001) in information theory and Markov models in probability theory. In contrast to AMs, this model enables the system to extract phrases of unlimited length and multiple morphosyntactic patterns. Counting the frequency of various consecutive tokens in the text is a simple and easily scalable statistical method that can yield large initial MWE candidates which then undergo further processing techniques.

Several MWE extraction researchers have adopted this model for retrieving various types of MWE (e.g., Choueka, 1988; Smadja, 1993a; Lin, 1998c; Gurrutxaga and Alegria, 2011; Frantzi and Ananiadou, 1996, among others). For instance, using the N-gram Statistics Package-NSP developed by Pedersen et al. (2011), Gurrutxaga and Alegria (2011) extracted various types of Basque noun-verb combinations by generating a bigrams list. They used two different window spans: ±1 and ±5, with a minimum frequency threshold of 30 per million words. The initial list of MWEs underwent several filtering stages to reduce data noise and the extracted inflectional words were normalised to their most common forms. In the final phase, the candidates were ranked based on several AMs and the result evaluated against a gold standard

list of noun-verb expressions. In another study, Silva and Lopes (2010) used the n-gram model to extract various types of MWEs representing fundamental concepts in the processed documents.

Several limitations of the n-gram model have been highlighted in the literature, such as the inadequate language modelling of discontinuous MWEs and the generation of large noisy data containing uninteresting sequences. However, MWE researchers implementing this model have proposed several tactics in the extraction process that can be used toeliminate the drawbacks mentioned above. For instance, Frantzi and Ananiadou (1996) and Smadja (993b) propose several methods and algorithms for enhancing the quality of n-grams when extracting nested or discontinuous MWE items. Also, combining the n-gram model with other statistical and linguistic techniques considerably improves the performance of this model in MWE extraction tasks. However, this is only a brief overview of the n-grams model; further details, including its main advantages and limitations, can be found in Manning and Schütze (1999).

2.2.2.2 Lexical Association Measures

MWE extraction methods based on statistical AMs are intensively used in the literature: the concept of AM is related to a distributional semantic hypothesis which assumes that lexical items with similar distributions usually have a similar meaning (Lin, 1998; Lin, 1999). Thus, the primary objective of most AMs is to statistically test the hypothesis that MWEs or collocations occur much more frequently than arbitrary consecutive tokens or any other combinations related to specific linguistic preferences. This form of hypothesis testing in AMs was illustrated by Seretan (2011,p. 35) who stated that:

> *'in testing word association, the alternative hypothesis is that the items* ***u*** *and* ***v*** *of a candidate pair are dependent on each other; the null (default) hypothesis is that there is no such dependence between the two items:*
>
> *– H 0 (null hypothesis):* ***u*** *and* ***v*** *are independent;*
>
> *– H 1 (alternative hypothesis):* ***u*** *and* ***v*** *are mutually dependent.*
>
> *The result of a test is given in terms of the null hypothesis, H 0: either H 0 is rejected in favour of H 1 (therefore, it can be concluded that H 1 may*

be correct), or H 0 is not rejected, which means that there was not enough evidence in favour of H 1'.

Many types of AMs have been used in research, and each has its advantages and limitations. For instance, in his collocation extraction experiment, Pecina (2005) lists more than 80 types of AMs used in the evaluation of multiple AM extraction models.

Several researchers adopt the use of the AM model in MWE extraction based on the significant frequency of co-occurrence tokens in the text (e.g., Church et al., 1991; Pecina, 2009; Moirón, 2005; Evert, 2005). However, no consensus was found regarding the preference for a specific AM score. Instead, research on comparative evaluations of AMs has shown considerable divergence when determining the best AM, which varies according to MWE type and the specific language domain.

AMs are usually limited to a restricted number of words in the collocation extraction model, which might make it difficult to adopt AM models when aiming to extract more extended sequences. However, to address this limitation, several researchers have attempted to modify or change the AM mathematical formulas to take account of more extended sequences. For instance, Mcinnes (2004) proposes an extension of the Log Likelihood ratio AM to discover MWEs that consist of more than two words. Another study by Moirón (2005a) extends the AM converge in extracting prepositional expressions by treating two or three tokens as one-string in the implementation of an AM extraction model. Table 2.3 shows a list of the common AMs that will be used as part of the AMWE extraction experiments reported in this thesis. More details on AMs have been presented in other research studies (e.g., Pecina, 2009; Moirón, 2005; Korkontzelos, 2010).

Table 2.3: Various common AM equations.

AMs	References	Formula
T-score	(Church et al., 1991)	$\frac{f_{xy} - \frac{f_y f_b}{N}}{\sqrt{\int xy}}$
Mutual Information (MI)	(Daille, 1994)	$log_2 \frac{f_{xy} N}{f_x f_y}$
MI3	(Daille, 1994)	$log_2 \frac{\int_{xy}^{3} N}{f_x f_y}$
MI.log_F	(Rychlý, 2008)	$MI - score \times log_{xy}$
logDice	(Rychlý, 2008)	$logDice = 14 + log_2 D = 14 + log_2 \frac{2fxy}{fx + fy}$
Log-likelihood(L.LK)	(Dunning, 1993)	$-2 \sum ij \int ij \ log \frac{f_{ij}}{f_{ij}}$

To identify a list of collocations from the corpus based on AM models, Ludeling and Kyto (2008) advise performing the following steps:

Choose an appropriate type of co-occurrence (surface, textual or syntactic).

Determine frequency signatures.

Filter the co-occurrence data set by applying a frequency threshold.

Calculate the expected frequencies of the word pairs.

Apply one of the simple AMs or produce multiple tables according to different measures (Ludeling and Kyto, 2008, p. 1242).

However, in the current research, different types of statistical models will be used to identify MWEs based on the adopted understanding of AMWE presented in this thesis. Thus, the extraction models will be based on a hybrid approach that utilises multiple statistical and linguistic components.

2.2.3 Hybrid approach

This approach is the most widely used in MWE extraction research because it takes advantage of knowledge-based and data-driven methods in the computational processing of MWEs. Linguistic and statistical approaches are considered complementary methods that enrich the effectiveness and quality of MWE extraction models. Utilising linguistic processing in statistical models results in more homogeneous and less noisy extraction findings. Thus, several studies have applied this methodology in MWE extraction to take advantage of the two language models and limit their weaknesses. For Arabic, Bounhas and Slimani (2009) presented a hybrid approach for AMWE extraction of compound nouns from a specialised corpus in the environmental domain. Figure 2.3 shows the architecture of the used extractor system, which consists primarily of three processing phases; morphological and POS analysis, a sequence identifier, and a statistical filter.

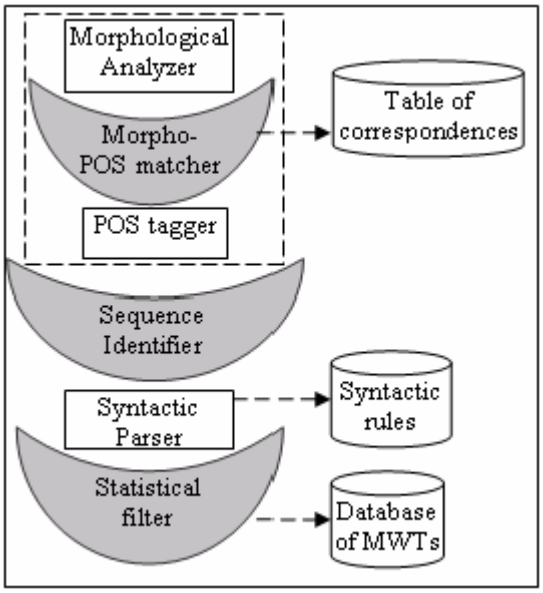


Figure 2.3: The Hybrid model for an environmental terms extractor

Another example of MWE acquisition, based on this approach, can be seen in Li and Lu's (2011) research which proposed a system framework for collocation extraction based on two modules; the bigram extractor and a synonym bigram noun phrase and verb phrases extractor. As shown in Figure 2.4, the collocation extraction goes through several phases in each module, followed by an evaluation of the extracted candidates which results in the final list of validated collocations.

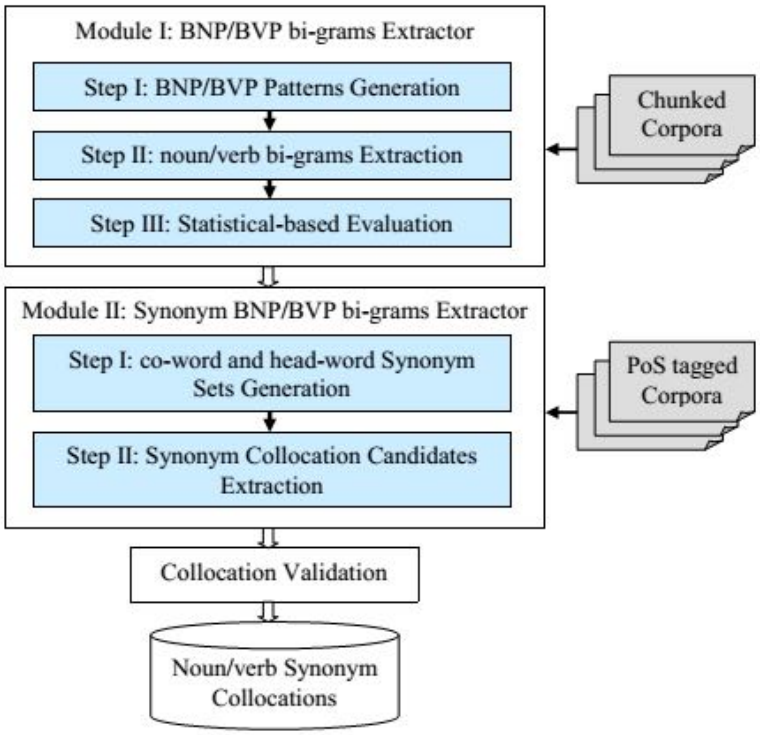


Figure 2.4: The hybrid framework for NP VP extractors (Li and Lu, 2011, p. 3).

Many studies emphasise the benefits of using multiple methods in MWE extraction due to the complexity of MWEs at various linguistic levels (e.g. Pecina, 2005; Seretan, 2011; Attia et al., 2010). In this thesis, a hybrid approach for AMWE extraction tasks will be applied; thus, linguistic and statistical methods will be implemented in an AMWE acquisition model to enhance the quality of its output.

2.3 Evaluation of MWE discovery models

Although considerable efforts have been dedicated to finding the best evaluation methods for the computational extraction tasks of MWEs (e.g., Evert and Krenn, 2001; Ramisch et al., 2012a; Thanopoulos et al., 2002; Krenn, 2008), no standard evaluation methodology has been proposed as the most appropriate for MWE extraction models. However, several evaluation methods have been developed for various experimental settings and according to the specific language domain (Ramisch et al., 2012). In the following subsections, existing evaluation methods identified in the literature will be briefly reviewed.

2.3.1 Expert judgments

This evaluation method has been applied in several MWE extraction studies, particularly in the absence of an appropriate gold standard evaluation of MWE LRs. It is based on the manual classifications of retrieved lists into true or false candidates or other ways of classifying positive outputs based on a specific annotation guideline that should be understood by all evaluators involved in the process.

Because this method requires manual work, the researcher should take into consideration recommended procedures such as training the annotators, providing them with clear guidelines, and measuring the inter-annotator score, which eliminates the claims of subjectivity commonly associated with manual annotation.

An example of how this evaluation method can be implemented was seen in research by Da Silva et al. (1999) who developed the LocalMaxs algorithm to extract contiguous and non-contiguous MWEs based on the use of various AMs. The evaluation task calculates the proportion of true MWEs in the extraction outputs to compare the performance of multiple AMs. In Seretan's (2011) research on the syntax-based extraction of MWE, this evaluation method is used to compare the outputs of syntax-based and window side statistical extraction models.

Many methods for measuring agreement among coders have been proposed in the literature, such as Cohen's κ (1960) or Fleiss' κ (1981). More details and an intensive survey of inter-annotator agreement can be found in Artstein and Poesio (2008) and Artstein (2017).

2.3.2 Comparison with existing MWE LRs

In this evaluation method, the extraction outputs are manually or automatically compared with constructed datasets by checking the candidates against available MWE LRs. Based on the classification finding which is reported in a matrix table, the precision, recall, and F measures are then computed for each MWE extraction task.

The evaluators who use this method assume that all candidates which do not match the evaluation LRs can be classified as false candidates. It is worth noting that, in several cases, especially when using short-coverage evaluation datasets, this method is used in conjunction with the manual annotation method described in section 2.3.1.

As mentioned previously, this method requires the use of existing evaluation MWE LRs; thus, this prerequisite condition limits the adoption of this method for languages with fewer or limited MWE evaluation LRs.

An example of MWE studies that use this method can be seen in the work of Riedl and Biemann (2015) who adopted the method to evaluate the outputs when using a distributional semantics model to rank domain-specific MWEs; an MWE annotated corpus which contains a list of annotated biomedical terms was used for the evaluation datasets.

2.3.3 Comparison with specially prepared gold standard datasets

Reference or gold standard data has long been used in the evaluation of various statistical methods in NLP and other related disciplines such as information retrieval. However, this method is also used frequently in MWE discovery experiments (e.g., Yazdani et al., 2015; Thanopoulos et al., 2002; Zilio et al., 2011). In these studies, multiple types of specially constructed MWE LRs were used as the reference datasets. For instance, Farahmand et al. (2015) developed an evaluation MWE LR that contains a list of 1048 MWE that were also classified into three categories based on their meaning:

Non-compositional.

Compositional but markedly conventionalised.

Compositional and non-conventionalised.

This evaluation dataset was then used by Yazdani et al. (2015) to evaluate multiple models predicting the non-compositionality of English MWEs.

The drawbacks of this method are the same as those of the previous method described in section 2.3.2 and relate mainly to the unavailability or the limited coverage of evaluation MWE LRs.

2.3.4 Task-based evaluation

This method is usually used to evaluate domain specific LRs when the constructed dataset aims to improve the performance of NLP tasks such as MT or semantic search. Thus, the primary goal of these methods in terms of evaluation is to measure the effect

of MWE LRs on the performance of NLP systems by comparing their performance before and after the integration of MWE LRs. As shown in many MWE research studies, the inclusion of MWE knowledge plays a critical role in improving the quality of many NLP tasks, such as MWE identification, language parsing, and MT (e.g., Costa-jussa et al., 2010; Villavicencio et al., 2007; Riedl and Biemann, 2016; Carpuat and Diab, 2010b).

2.4 Research timeline for related MWE language resources and their computational representations

In this section, a survey of existing diverse MWE repositories will be presented with a focus on AMWE LRs. Projects in this area have attempted to create an electronic database for multiple types of MWEs that cover various morphosyntactic structures and semantic domains. The SIGLEX-MWE website lists more than 22 MWE resources in different languages; these are open source projects available for download.[16]

In this regard, it is important to note efforts towards parsing and multiword expressions within a European multilingual network (PARSEME), which is an on-going project involving a multidisciplinary research community devoted to studying MWE phenomena in multiple European languages, especially in relation to language parsing and linguistic resources (Savary et al., 2015). As part of their research in this area, Losnegaard et al. (2016) conducted a survey on available MWEs LRs based on the result of an online questionnaire which was designed to obtain detailed information on existing MWE resources. The survey used an online form as the crowdsourcing tool for information on MWE LRs. The form was divided into two main sections. The first was devoted to questions eliciting general information about LRs such as language name, type, size, online link, and so on. The second section aimed to obtain more detailed information about the LR, such as relevant publication, annotation schema, and grammatical and lexical frameworks.

The core aim of the survey was to provide the end user with an overview of the most available MWE LRs with all the necessary details about their development and

[16] http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets

accessibility. The survey showed that although there are many MWE LRs available, detailed information about them is scarce and difficult to find. This is especially the case for non-European languages (such as SA) and for LRs not registered in public international LRs infrastructures, such as the following catalogues:

META-SHARE: the ILSP managing node.

ELRA: European Language Resources Association.

SIGLEX-MWE: the MWE community website.

The survey results are publicly accessible as an online updated spreadsheet.[17] Based on the main classifications of the study questionnaire,[18] the MWE LRs were grouped into five categories, as shown in Table 2.4.

Table 2.4: The main types of available MWE LRs.

Nu	MWE LRs	LRs Count	Percentage
1	Treebank with MWE annotations	12	11%
2	MWE lexicons	48	45%
3	Monolingual list of MWEs	13	12%
4	Multilingual resources	15	14%
5	Others (for all the LRs not in the previous categories)	19	18%
	Total	107	

Regarding the length of MWE lexical entry in these LRs, the range is from only two to 23 lexical components which includes adjacent structures and other more flexible and adjustable MWEs. The size of these LR ranges from a few hundred MWE items with different types and layers of linguistic annotation to a large LR which exceeds three hundred thousand, mostly plain, MWE entries.

Regarding the public accessibility of the LRs, the researchers found that 40 out of 107 resources were freely available for researchers under the creative commons licence.[19] However, not surprisingly, English is the dominant language in all these MWE LRs, although other European languages can also be found such as German, Croatian, Greek, and Portuguese. The findings show that the vast majority of the

[17] https://sites.google.com/site/mwesurveytest/home.

[18] The online survey form: `https://goo.gl/eYz8qL.`

[19] More details about this licence can be accessed through this link:
`https://creativecommons.org/licenses`

MWE LRs developed were devoted to NLP applications while several others were meant for human users, such as the language learner LRs.

In the following subsections, a timeline research review of related works on developing multiple MWE LRs and computational lexical representations will be presented. The focus will be on LRs not included in the popular online linguistic databases and on AMWE LRs more relevant to the research questions and objectives of this thesis.

2.4.1 A Database of Lexical Collocations (Krenn, 2000a).

Krenn (2000a) built a lexicon of German prepositional collocations which consisted of one thousand items. The phrases in this LR were represented in a relational database model that includes various types of linguistic description. Data collection was based on the use of manual methods from traditional dictionaries and on the use of a statistical model for extracting corpus-based instances for targeted collocation patterns. The representational model consists of four main relations: collocation-instance, ci-analysis, collocation-realisation, and cr-structure. The first two relations represent the competence base while the others represent the example base. Every relation has a list of attributes that provide information about the linguistic features of the collocations. However, the relation model exhibits several limitations when representing various types of linguistic information. This is because the morphosyntactic knowledge tends to be complicated and changeable, especially in morphologically rich languages. Thus, Extensible Mark-up Language (XML) representations provide an alternative and more reliable and flexible way of representing lexical LRs.

This LR was updated and later extended to include 21796 German combinations of prepositional phrases (Krenn, 2008). Table 2.5 presents basic statistical information about the lexicon.

Table 2.5: Basic information about the German prepositional phrase LR.

Type	Number	Percentage
True positive collocations	1149	5.3%
Verb-object collocations	549	2.5%
Figurative expressions	600	2.8%
Collocations found in. french corpus 30	5102	23.4%
Light verbs	6892	31.6%
total	21796	100.0%

The representation model in this study focused mainly on syntactic information which meant that other levels of linguistic analysis were absent in the lexicon model. However, the classifications presented in this study could be applied to several types of AMWE and the type of linguistic description can be adapted to lexical entries with several modifications to align with the linguistic properties of SA.

2.4.2 A Scientific Arabic Terms Database (Lelubre, 2001).

This research represents an early attempt to build a phraseological list of terms in SA whereby a domain-specific lexicon of scientific terms in the field of optics was developed with translated versions in French and English. The database consisted of 6k terms and was collected manually from related corpora, journal, and SA handbooks on physics. The syntactic structure of terms ranges from single-word to multiple types of compound MWE terms containing more than one word.

Three classifications were adopted in this project, as shown in Figure 2.5. The primary field of the lexicon are the terminological units which include the components of the lexical entries, the terminographic data, and referential fields containing additional information about the terms (e.g., synonym, abbreviation, definition, and original LR).

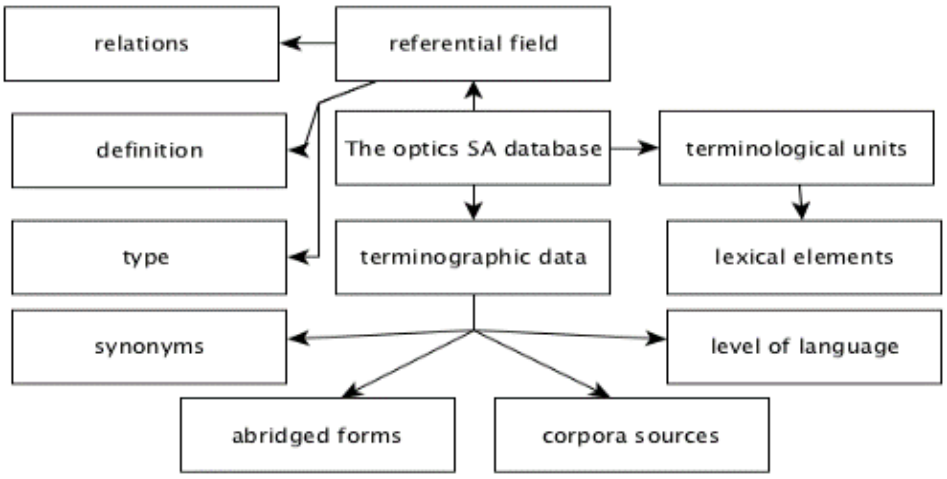


Figure 2.5: Classifications included in the optics terminological databases (Lelubre, 2001).

In the representational relational model, the focus was on the morphosyntactic features of elements of the term based on several features or specifiers. Table 2.6 presents examples from the information included for the constituents of MWEs in this LR.

Table 2.6: Linguistic features included in the AMWT lexicon of the optics (Lelubre, 2001).

Linguistic features	Code
number of plural forms	Nb Pl
number of type of declension	td
gender	g
number	n
kind of determination	d

Although scientific MWTs will be excluded in the lexicon developed for this thesis, as will be described in Chapter 3, Lelubre presents a list of linguistic properties of several syntactic structures in SA that might also be shared with other AMWE included in the context of the current research.

2.4.3 Word frequencies in written and spoken English (Leech et al., 2001).

In terms of the English language, Leech et al.'s (2001) work on the development of the 100-million word British National Corpus (BNC) (Leech, 1993) is considered to

be one of the earliest attempts to construct a corpus-informed phrase list. Table 2.7 presents several examples of the phrases included in this LR.

Table 2.7: Sample from the phrase list of Leech et al. (2001).

Word	POS	Derivations	Frequency (p/million)
A bit	Adv	:	119
A great deal	Adv	:	14
A little	Adv	:	104
A lot	Adv	:	40
Abandon	Verb		44
		Abandon	12
		Abandoned	26
		Abandoning	5
		Abandons	1
Abbey	NoC		20

The generation of this list was based on the automatic extraction of the most common phrases appearing in the POS-tagged written and spoken corpus. The entire BNC was run through the Constituent-Likelihood Automatic Word-tagging System (CLAWS) (Garside, 1987) which classifies words into their morphological categories (e.g. noun, verb, adjective, adverb).

The criterion for MWE selection adopted in the research was phrases with a high degree of fixedness or non-compositionality. MWE were defined in this study as ‘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, p. 8).

Based on the adopted specifications of MWEs, several other types of MWE were excluded, such as syntactically flexible expressions or discontinuous MWEs in the form of phrasal verbs (e.g. write down, write it down). However, this research was a list for English MWEs that provides rich language data on frequently used words and fixed phrases in the written and spoken BNC. In the current study, fixed expressions were included in extraction models and various tactics were adopted to also include flexible and discontinuous AMWEs.

2.4.4 Representational model for MWE Lexicons (Calzolari et al., 2002).

This study reflects Calzolari et al.'s (2002) effort to construct a uniform multi-lingual MWE representational model which includes syntactic and semantic information in XML form. According to the researchers, A MWE is 'a sequence of words that act as a single unit at some level of linguistic analysis' (ibid, p. 1934). Furthermore, although they admit some difficulty in specifying precise boundaries for MWEs, they propose a set of criteria for defining MWEs that includes the following linguistic properties which should be considered when discovering MWE knowledge:

reduced syntactic and semantic transparency;

reduced or a lack of compositionality;

more or less frozen or fixed status;

possible violation of some otherwise general syntactic patterns or rules;

a high degree of lexicalisation (depending on pragmatic factors);

a high degree of conventionality' (ibid, p. 1934).

The researchers build on previous work aimed at constructing a model for lexical information (Romary et al., 2000) and extend the model to accommodate multi-layered encoding of MWE knowledge. The representational model proposes the inclusion of most types of linguistic information and considers their potential variants. The model was then refined and reviewed later in Francopoulo (2013) as part of the lexical mark-up framework (LMF) project which aimed to establish standards for representing multiple types of LR in computational forms, as will be reviewed in section 2.4.20. In the current study, these previous efforts will be taken into account when constructing a lexicon model for AMWEs.

2.4.5 A syntactically annotated idiom dataset (Kuiper et al., 2003).

This study involved building an idiomatic expressions dataset to explore the syntactic behaviour of these phrases in various linguistic contexts. One of the central hypotheses this study investigated was the idiosyncratic nature of these types of phrases in English at multiple levels of linguistic analysis. Fixed and flexible

expressions that represent most structural types were included in this study to provide a comprehensive dataset that reflects the actual use of idioms in English. The lexicon comprises 13,467 phrasal lexical items (PLIs), and the collection methods involved extracting all the data from four previously published dictionaries of English idioms.

The authors targeted English linguists and second language learners as the end-users of this computational LR. The final lexicon was presented in a txt file and, based on the generative framework, the data were manually analysed. Three main reasons were used to justify the use of manual methods, which are as follows:

'First when the analysis began …, machine parsers were not able to provide sufficient detail. Second, manual annotation raised questions about the best analysis which was heuristically challenging. Third, the period taken for the analysis allowed many people to work on the project both with analysis and checking and this has led to a perhaps more considered analysis than what might have been done with faster machine parsing'(Kuiper et al., 2003, pp. 4-5). Table 2.8 presents examples of the conventions adopted in their manual analysis that were added to the original txt file.

Table 2.8: Examples of symbols used in the analysis of English idioms.

Conventions	Functions
[]	enclose constituents.
/	is placed between alternative heads (selection sets).
()	is placed around lexicalised optional constituents.
*	indicates an ungrammatical PLI
NP	is used for many slots.

The AMWE LR in the current study will not adopt any specific linguistic theory in the analysis. Instead, the AMWE will be presented in multiple formats which provide linguistically useful data that can be applied in a wide range of existing linguistic frameworks. In the extraction and collection processes, hybrid corpus-based methods will be adopted to ensure the representation of the actual AMWE used rather than an existing written dictionary primarily constructed through traditional and older manual methods of collection.

2.4.6 Collocation and synonymy in classical Arabic (Elewa, 2004).

This research aimed to implement a corpus-based analysis of collocations in CA, especially those with semantic relations such as the synonymy or non-synonymy of

these lexical units. The definition of collocation adopted was based on the main characteristics of this phenomenon mentioned in the literature; thus, in this study, collocation include the co-occurrence of at least two adjacent or non-adjacent words without any syntactic restrictions (Elewa, 2004, p. 54).

Although there was no intention in this study to develop an AMWE LR, the corpus-based analysis of extracted collocations provided invaluable information about the syntactic and semantic behaviour of collocations in CA. This is also a useful resource for studying AMWEs in SA, which is the intention in this thesis.

The extraction methods involved a range of statistical and linguistic components based on a list of randomly selected synonyms that were used as node words in extracting relevant collocations. A new classical Arabic corpus contained 5 million words collected by the researcher that represent four main semantic genres: thought and belief, literature, linguistics, and science. Table 2.9 presents examples of the node words used in this study.

Table 2.9: Examples of synonym words used in studying the semantic relations of CA.

POS	set of synonyms		
V	جاء / أتى	jāʾ / ʾatā	come
V	ظن / حسب	ẓann / ḥasib	think
N	إثم / ذنب	ʾiṯm / ḏanb	sin
N	ود / حب	wadd / ḥubb	love

The typology of phrases adopted in this study was based on the degree of flexibility of the expressions, as shown in Figure 2.6 with Arabic examples.

Figure 2.6: Collocation typology adopted by Elewa (2004).

2.4.7 An automatically built Named Entity lexicon (Attia et al., 2005).

Using Arabic wordnet and Arabic Wikipedia, Attia et al. built a large LR of 45,000 Named Entities (NEs). The extraction methods were based on four main processing

phases: mapping, extraction, postprocessing, and diacritisation. First, 1572 instantiated synsets were extracted from the selected LRs, and then part of the extracted nouns were mapped to the relevant categories in the wordnet. Second, the related articles were retrieved from Wikipedia and, using inter-lingual links, keyword searching, and regular patterns of expression, lists of items related to only two types of NEs were identified: person and location, as these were high in frequency. In the extraction process, a list of over 60 keywords was employed to enhance the automatic extraction of numerous similar NEs. Table 2.10 provides several examples of these words.

Table 2.10: Examples of keywords used in the automatic extraction of NEs (Attia et al., 2005b, p. 3616).

Search keywords	Examples of extracted NEs		
دولة	دولة الصين	dawlat aṣṣīn	State of China
قرية	قرية النهر	qaryat annahr	River Village
بحر	بحر العرب	baḥr alʿarab	Arabian Sea
جبل	جبل أحد	jabl ʾuḥud	Mount Uhud
مدينة	مدينة جدة	madīnat jiddah	Jeddah city

Third, the post-processing phase target was to extend the extracted list of NEs by implementing several mappings and comparisons between multilingual LRs. Thus, in this step, NEs found in other languages were considered potential NEs in Arabic and vice versa. The final processing phase in this study involved adding diacritics to the extracted NEs that play a prime role in eliminating language ambiguity in various NLP tasks. To achieve this, a unique diacritisation pipeline was developed which utilised both linguistic and statistical methods, as shown in Figure 2.7. The system output led to the discretising of 73% of the NEs.

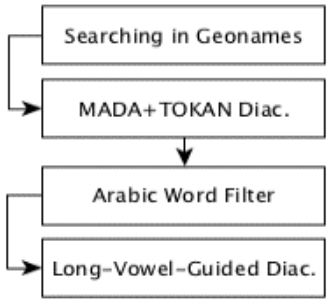


Figure 2.7: Arabic NEs diacritisation pipeline used by Attia et al. (2005b, p. 3617).

Although this study does not share the concept of AMWE adopted in the current research, the large repository of NEs it built provides valuable information about the advantages of using a hybrid model in the extraction process and explores various structural types of NEs that are similar to the types of AMWE included in the current study.

2.4.8 Comparing and combining a semantic tagger and a statistical tool for MWE extraction (Piao et al., 2005).

This study applied an automatic hybrid model to extract MWE using the English semantic tagger (USAS) (Rayson et al., 2004).[20] This was developed at Lancaster University based on POS annotation provided by the CLAWS tagger (Garside and Rayson, 1997). The research built on previous work reported in Piao et al. (2003). The extraction experiment was based on court issues in newspaper sections of the Meter corpus (Gaizauskas et al., 2001) which consists of 774 articles and 250,000 words. The study adopted the broad practical definition of MWEs proposed by Biber et al. (2003) which describes MWE as 'combinations of words that can be repeated frequently and tend to be used frequently by many different speakers/writers within a register' (Piao et al., 2003, p. 53). The extraction model used symbolic and statistical tactics to enhance the final outputs and take advantage of the MWE template lexicon which was developed as part of the USAS and contains over 18,600 MWEs.

The extraction used the USAS system to tag the corpus with POS and semantic tags. The MWE assigned by the tagger as one semantic unit was then collected and manually evaluated. The initial list of candidates consisted of 4195 items, which was later reduced to 3792 after manual evaluation. The authors reported a precision score of 90.39%. The recall score was estimated to be 39.38% based on the evaluation of a sample from the corpus annotated with MWEs due to the unavailability of a fully MWE annotated corpus.

Semantic analysis of the findings shows that MWEs in English belong to most of the semantic fields used by the USAS tagger and the majority of retrieved items were

[20] More information about the English semantic tagger can be accessed through http://ucrel.lancs.ac.uk/usas.

semantically classified as names and grammatical words with 1635 MWE candidates[21].

Furthermore, the study provides evidence that most MWE candidates can be found in sequences from two to four words with the dominant extraction of bigram MWEs that constitute 3,105 true items of the data. In the AMWE extraction used in the current research, a length restriction of two to six words will be employed and this does not include discontinuous AMWE candidates which might consist of more than six words. Piao et al. (2005) also report a contrasting relationship between the frequency of MWE and the extraction precision scores in that the high frequency items yield lower precision scores and vice versa. Hence, the current research on AMWE will explore the estimated performance of an extraction model within various AMWE lengths and at various levels of frequency.

In the second part of the research, statistical methods will be used to extend the coverage of symbolic tools in extracting MWEs. The model used includes the implementation of five phases, which are as follows:

(1) POS-tag the input text using a CLAWS POS tagger.

(2) Collect collocates using the co-occurrence association score.

(3) Using the collection of collocates as a statistical dictionary, check the affinity between closely adjacent words to create an affinity distribution map.

(4) Based on the affinity distribution, collect word clusters (not just word pairs) that are subject to relatively stronger affinity.

(5) As an option, apply simple filters to clear highly frequent errors (Piao et al., 2005)

These processing steps led to the extraction of 3306 candidates with a precision score of 81.85%. The following example is an output text that represents samples of MWEs annotated with POS and also the tag pairs <mwe> </mwe>:

[21] The USAS uses 232 semantic field categories which are grouped into 21 general discourse fields.

<s><mwe> Deputy_NN1 principal_NN1</mwe> Alden_NN1 was_VBDZ jail-ed_VVN for_IF15_MC years_NNT2 after_II being_VBG<mwe> found_VVN guil-ty_JJ</mwe> of_IO five_MC <mwe> indecent_JJ assaults_NN2</mwe>,_,one_MC1 gross_NNO indecency_NN1 and_CC four_MC <mwe> serious_JJ sexual_JJ assaults_NN2</mwe>._.</s>

Comparison of the findings between the symbolic and statistical methods in MWE extraction reveal that the integration of these methods substantially improves the coverage of the extraction model and in enhances the abilities of the model to extract multiple types of domain-specific MWEs.

2.4.9 Semantic lexicons for corpus annotation (Piao et al., 2006).

This study reports the development of large-scale general-purpose semantic lexicons that have been built at Lancaster University for more than 14 years. The lexicons were manually constructed by linguists and consist of 45,800 single word entries and over 18,700 MWE template entries. These are semantically annotated with 232 semantic tags classified under 21 main semantic field categories, as shown in 2.11

Moreover, the lexical entries in each semantic field are divided into multiple categories based on their sub-type meaning or semantic relationships, such as synonym-antonym or meronymy-holonymy.

For instance, the food and farming category includes multiple types of sub-meaning fields such as drink and drugs. The lexicons adopt the use of a list of codes that represent all the semantic fields and different types of semantic relationships (e.g., A15+ = Safe. and A15- = Danger). Table 2.12 presents a list of MWE instances along with their semantic labels from the MWE lexicon.

Table 2.11: The 21 major semantic fields of Lancaster.

Code	Semantic fields	Code	Semantic fields
A	General and abstract terms	B	The body and the individual
C	Arts and crafts	E	Emotion
F	Food and farming	G	Government and the public domain
H	Architecture, buildings, houses, and the home	I	Money and commerce in industry
K	Entertainment, sports and games	L	Life and living things
M	Movement, location, travel, and transport	N	Numbers and measurement
O	Substances, materials, objects, and equipment	P	Education
Q	Linguistic actions, states, and processes	S	Social actions, states, and processes
T	Time	W	The world and our environment
X	Psychological actions, states, and processes	Y	Science and technology
Z	Names and grammatical words		

The distribution of MWEs in these semantic categories shows that names and grammatical words were the dominant fields with 3,137 items, followed by general and abstract terms with 2,160 MWEs. The total number of semantic types found in the MWE lexicon was 2,763 tags representing various semantic domains.

Table 2.12: Examples of MWE lexical entries with semantic annotation.

MWE templates	semantic annotation
Child*_NN*Protection_NN1 Agency_NN*	Z3c
take*_* {Np/P*/R*} for_IF granted_*	T3/X2.6
life_NN1 expectancy_NN1	S1.2.3+

The semantic MWE LR developed in this study can be used in multiple practical applications, such as semantic tagging, automatic word classification, and the extraction of new MWEs in various semantic domains. In the proposed lexicon the intention is to collect AMWEs from multiple semantic domains. Therefore the semantic taxonomy of Lancaster will be adopted in the semantic representations of AMWEs.

2.4.10 A multilingual collocation dictionary (Cardey et al., 2006).

In this project, a multilingual collocation lexicon was developed for translation purposes that covers various language domains. The dictionary contains multiple

types of phrases that range from wholly fixed expressions to more flexible and uncontentious phrases at different non-compositionality semantic levels (e.g., kick the bucket, medical history, spill the beans). In the lexicon model, the first words of the collocations are used as the headword in the dictionary. The model was based on a previously developed collocation system which includes four main elements for representing each lexical entry as follows:

The headword of the collocation associated with its synonyms, translations, and polysemic equivalences.

List of collocations related to the headword.

Sense group for each collocation across several languages.

The language ID for each lexical entry.

The grammatical category and function of collocations.

The AMWE entries included in this dictionary are very limited in terms of their size, and so the dictionary was primarily used as an additional translated version of the English collocation lexicon. In the current study, several linguistic features of the lexical representations adopted in this research were used to enhance the usability and scalability of the AMWE LR.

2.4.11 German idioms and light verbs (Fellbaum et al., 2006).

In this project, a large lexical database was built for German verb phrases, idioms, and light verbs to reflects the usage of these types of expression based on corpus-based evidence.

The corpus used in Fellbaum et al's (2006) extraction process comprised over a billion words of German newspapers representing various language genres. The extraction process was mainly based on the extraction of collocates related to the most frequent verbs and nouns in German. The results were then checked manually by lexicographers and linguists and, sometimes, based on this validation the initial corpus query was refined and improved in the second round of MWE extraction.

An example corpus was created which contains numerous examples of sentences containing idioms and phrasal verbs in various linguistic contexts. These enhance the usability of this LR by linguists in exploring MWE behaviour at multiple levels of

analysis. Concerning the representations, the MWEs extracted in this study improve with various, comprehensive types of linguistic metadata, as shown in Table 2.13.

Table 2.13: The representations of German MWEs (Fellbaum et al., 2006, p. 358).

Levels	Types of annotation
1	the citation form
	corpus occurrences,
	information about usage and alternations
	polysemous and homonymous idioms
	additional free-form comments
2	a dependency structure to represent the phrase structure
3	morphosyntactic properties
4	lexical and phrasal variations
5	the syntactic transformations found in the example corpus of each idiom
6	semantic features (e.g., semantic field and a domain label)
7	paradigmatic relations among the idiom under consideration and other idioms.
8	information about the example corpus (e.g., name, source and the search queries),
	the idiom's template, and various administrative options.

The types of information added to each lexical entry of this LR provide a valuable resource for exploring all the related linguistic phenomena of the expressions under consideration. In this thesis, comprehensive representations of the targeted AMWEs will be adopted that include most of the metadata in Felbaum et al.'s research.

2.4.12 Arabic multi-word expressions datasets (Attia, 2008).

In this study, a set of AMWE lists were collected as part of the process of developing an Arabic morphological and syntactic disambiguation system using the LFG framework. This framework was based on the Xerox Linguistic platform created by Butt (1999; and Dipper et al. (2004) for writing language grammar rules and carrying out various linguistic levels of analysis.

The AMWE items were extracted using a corpus concordance tool as well as by using manual collection methods. The AMWE transducer built in this study was used as a complement to the morphological transducer which aims to handle the language ambiguity caused by multiple types of AMWEs.

When defining AMWEs, Attia followed the criteria specified previously by several researchers (e.g., Baldwin and Tanaka, 2004; Calzolari et al., 2002; Guenthner and

Blanco, 2004). These criteria can be summarised in terms of the following properties for identifying MWE:

Fixedness of the phrase. This feature can be displayed in different ways, for instance by replaceability, so that the word ‘many' as in ‘many thanks' cannot be replaced by similar adjectives such as ‘several'.

Non-compositionality of the phrase. This semantic feature means that the meaning of the phrase is not obtained from its component parts (e.g., kick the bucket = die).

Syntactic irregularity. The phrase has a particular syntactic form which is different from regular syntactic structures (e.g., by and large)

Single-word replacement. One single word could replace the phrase (e.g., give up = abandon, looking glass = mirror).

Translation. This can be used to identify MWEs when we can see the corresponding phrase or word in other languages (e.g., looking glass = مرآة mir’āh (in Arabic) (Attia, 2006, p. 88).

Based on the classifications of MWEs presented by Sag et al. (2002), Attia classified AMWEs into five categories according to semantic compositionality and syntactic flexibility. Table 2.14 presents examples of AMWE constructs included in this study with examples.

Table 2.14: Types of AMWEs with examples (Attia, 2008, pp. 79-84).

AMWE types	Examples		
Fixed compound nouns	حفظ الأمن	ḥifḏ alʾamn	Peace-keeping
Semi-fixed expressions	قصير النظر	qaṣīr annaḏar	short-sighted
Linking expressions	وعلى هذا	waʿalā hāḏā	whereupon
Prepositional expressions	بشكل جذري	bišakl jaḏrī	fundamentally
One string expressions	بالتالي	bittālī	consequently

However, there was no intention to create detailed lexical representations for AMWEs in Attia’s study because the sole aim was to improve the morphological analyser system by accommodating a list of AMWEs. The findings emphasise the significant role of MWE LRs in improving the system’s output.

2.4.13 An Arabic Multiword Term Extraction Program (Boulaknadel et al., 2008).

Boulaknadel et al. (2008) designed and developed an AMWTs extraction program. The research was applied to a 475,148 word specialised corpus related to the environmental domain. A reference list made up of 65k MWTs was also created to automatically annotate the findings of the statistical measures. The MWT extraction was based on a hybrid approach that takes advantage of linguistic specifications for detecting AMWTs and also existing statistical AM models.

Regarding the linguistic specifications, the authors listed several properties of Arabic MWTs that should be taken into consideration during the identification and extraction of Arabic terms. For instance, regarding the morphosyntactic structures of expressions, they found that most AMWTs belong to the familiar morphosyntactic patterns found in English and other languages (e.g., N-ADJ, N-N). The authors also considered linguistic variations in MWTs that affect the extraction process. The Arabic language exhibits vibrant inflectional variation including several types of noun inflection such as number and gender as well as adjectives, and the definite article which appears intensively in AMWTs. Hence, consideration of all these linguistic parameters plays a significant role in the improvement of precision when identifying and extracting these types of phrases.

To measure the association strength for the extracted MWTs, the researchers used four types of AM algorithms: LLR (Dunning, 1993), T-score (Church et al., 1991), FLR (Nakagawa and Mori, 2003), and Mutual Information (MI^3) (Daille, 1994)). They then conducted a comparative evaluation of the AMs to discover the best performing algorithms for extracting AMWTs in the environmental domain. The findings presented in Table 2.15 show that the LLR outperforms other AMs with a precision score of 85%, followed by FLR with 60%.

Table 2.15: The precision scores of AMs in extracting MWTs.

Type	P(%)	Type	P(%)
FLR	60%	LLR	85%
T-score	57%	MI3	26%

Because the current research will adopt the use of a hybrid model in the extraction process for multiple types of AMWEs, the evaluation procedures of Boulaknadel et

al's (2008) study can be implemented in the thesis. In addition, the results they obtained provide insights into the best performing AMs for extracting MWTs in the environmental domain.

2.4.14 Arabic multi-word term extraction (Bounhas and Slimani, 2009).

This research attempted to implement other extraction techniques for compound nouns which are generally based on hybrid models. Moreover, they proposed new algorithms to reduce morphological and syntactic ambiguities during the extraction process. Their model consists of three phases starting from the morphological analysis which is followed by the sequence identifier and syntactic parser; the final results are then filtered based on the statistical information. The extracted items were classified into six categories according to different types of Arabic Compound noun, as shown in Table 2.16.

Table 2.16: Classifications of nominal MWEs with examples (Bounhas and Slimani, 2009).

Nu	AMWE classifications	Examples		
1	Annexation	the car of a wealthy man	سيارة الرجل الغني	sayyārat arrajul alġanī
2	Adjective	a rich man	رجل غني	rajul ġanī
3	Substitution	this car	هذه السيارة	hāḏihi assayyāra
4	Prepositional	a kind of sweet	نوع من الحلوى	nawʿ min alḥalwa
5	Conjunctive	the cat and the mouse	القط و الفأر	alqiṭṭ wa alfaʾr
6	Compound nouns linked by composite relations	To persist for about one year	الاستمرار لحوالي سنة	alistimrār liḥawālay sana

The final list of AMWEs ws compared to a previously developed MWE list, and the evaluation shows improvements in extraction accuracy over previous experiments applied to MWE acquisition in the same language domain.

2.4.15 Dutch Multiword Expressions lexicon (Grégoire, 2009).

This project constructed a Dutch electronic lexicon of MWEs to improve the treatment of MWE in the task of identifying and enhancing various Dutch NLP systems. The extraction of MWEs was based on automatic methods from several corpora that underwent manual evaluation for inclusion in the lexicon. The extraction process was based on an analysed corpus by Alpino, which is a Dutch-specific language parser.

The model used six predefined syntactic patterns in the extraction tasks, as shown in Table 2.17. The selection of these patterns was based on the frequency information of the corpus where, for each candidate, related information was extracted such as the subcategorisation frame, a list of heads of co-occurring subjects, and number information of the noun.

Table 2.17: Basic information about the MWE extraction process (Grégoire, 2009).

Selection pattern	Number of extracted candidates
NP_V	3,894
(NP)_PP_V	2,405
NP_NP_V	202
A_N	1,001
N_PP	1,342
P_N_P	607
total	9,451

The representational model for MWEs adopted the equivalence class method (ECM) used by Odijk (2003) which was extended to accommodate MWE knowledge. The representations were divided into two sections, MWE pattern description and MWE description; the former was devoted to describing the core properties of the MWE pattern while the latter describes the related linguistic information for individual MWEs. Table 2.18 presents examples of the features included in the lexicon representational model.

Table 2.18: Examples of the MWE description included in Grégoire (2009, p. 41).

Features	Example
PATTERN _ NAME	ec1
POS	dnv
PATTERN	[.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]
MAPPING	345
EXAMPLE _ MWE	de boot missen
EXAMPLE _ SENT.	hij heeft de boot gemist
DESCRIPTION	Expressions headed by a verb, taking a direct object consisting of a fixed determiner and an unmodifiable noun.
COMMENTS	

However, in the extraction of AMWEs, the current research will use predefined selection morphosyntactic patterns. Thus, it is helpful to investigate whether the patterns in Grégoire's study are common in SA. Additionally, the representational

model constructed for this research provides informative insights into the most critical linguistic descriptions and these should be included to enhance the use of AMWE LR in the current study.

2.4.16 Automatic extraction of Arabic multiword expressions (Attia et al., 2010).

This research investigates the automatic extraction and acquisition of AMWEs from multiple LRs. The study was based on three data resources used for compiling MWEs (Wikipedia, Princeton WordNet, and Arabic Gigaword [Fourth Edition]). The complex nature of extracting and identifying MWEs meant the authors depended on three different approaches to the extraction and evaluation of AMWEs based on the availability of rich language data:

Cross-lingual correspondence asymmetries.

Translation-based extraction.

Corpus-based statistics.

The study focuses on nominal MWEs, justified by the statistical data contained in WordNet, which shows that the most frequent AMWEs are compound nouns. The first model was created to capture non-compositional MWEs items based on the translation of MWEs items in different languages. The core assumption of this method is that an MWE item has no mirrored representation in other languages or it can be translated into a single word.

Thus, to discover idiomatic AMWEs this model followed three main steps:

A candidate selection which included all the Arabic Wikipedia MWE titles.

A filtering process that aimed to exclude ambiguous titles and administrative pages.

A validation process that compared the MWE candidates with their translations in several languages and the corresponding WordNets.

This technique was based on the previous study conducted by Zarrieß and Kuhn (2009) for extracting nominal types of MWEs. When the translation in any language is a single word, it is classified as a potential AMWE. An example of this method can be seen in the Arabic phrases " فقر الدم *faqr addam"* which translates into the English

single word ‘anemia’. The second approach focused on identifying compositional compound nouns by extracting a list of MWEs in WordNet and looking for their equivalent translation in Arabic; it assumes the English MWEs are likely to be considered AMWEs. In the next step, various search queries were utilised to filter the translation results. This technique also builds on previous research by Vintar et al. (2008) albeit with several modifications to the translation methods. For example, instead of using an alignment-based approach, the researchers used an MT system. The final approach used AMs to extract MWEs from an 848 million-word unannotated Arabic corpus based on the frequency distribution of co-occurrences. The conclusive findings of this research yielded substantial lists of Arabic MWEs. Table 2.19 presents a comparison of the size of AMWE lists based on each approach.

Table 2.19: The size of AMWEs lists based on each approach (Attia et al., 2010b, p. 26).

Extraction method	MWEs	Intersection
Cross-lingual	7.792	------
Translation-based	13,656	2658
Corpus-based	15,000	697

Several reasons are given to justify the low level of intersection among the findings of the three MWE extraction approaches. For instance, it might be due to the different nature of the adopted LRs: Attia et al. (2010) indicate that many MWEs extracted using AMs from the Gigaword corpus do not have equivalents in SA LRs such as Wikipedia and WordNet. Table 2.20 presents examples of AMWEs retrieved in this research using multiple discovery methods.

Table 2.20: Examples of AMWEs extracted by Attia et al. (2010b).

AMWE			LR	method
Craterellus	فطريات دعامية	fiṭriyyāt daʿāmiyya	Arabic Wikipedia	Cross-lingual
Cockpit	قمرة القيادة	qumrat alqiyāda		
Jellyfish	قنديل البحر	qindīl albaḥr		
Two memorabilia	درعـــيـــن تذكاريتين	daryn tḏkāriyyatayn		
Shyam Saran	شيام ساران	šayām sārān	Arabic Gigaword	AMs
Haafat Maon	هافات ماعون	hāfāt māʿūn		
Life	حياة	ḥayā	English Princeton WordNet	Translation-based
Eye contact	التقاء العيون	iltiqāʾ alʿuyūn		
Market penetration	اختراق السوق	iḫtirāq assuq		

The findings present candidates for AMWE sthat cannot be found in most written published LRs. This emphasises the idea that MWE knowledge in SA is rich and changeable over time, thus new methods and research are urgently needed to extract new AMWE items .

2.4.17 Multiword expressions and named entities in the Wiki50 (Vincze et al., 2011).

This study annotated multiple types of MWE and NE candidates in English Wikipedia to use them as training datasets in MWE and NE identifier systems. The data source (Wiki50) contains 50 articles which include at least 1k words excluding structured texts (e.g., lists, tables and figures). After manual segmentation of the corpus, the researchers identified 4350 sentences in the data. They define MWEs as ‘lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy' (ibid, p.289). The NEs covered in their annotation include four main classes: persons, organisations, locations, and miscellaneous. Figure 2.8 presents the types of MWE included in this study with examples.

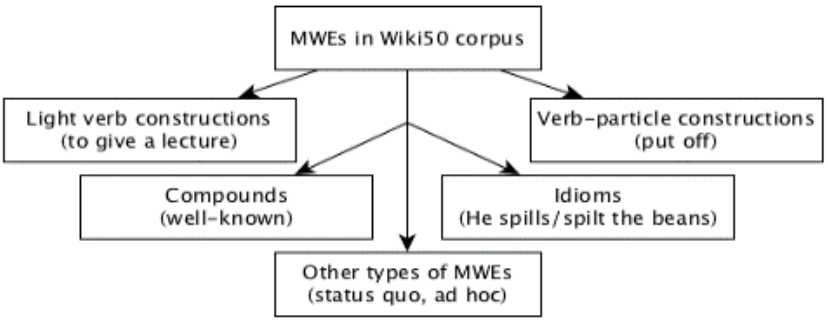


Figure 2.8: MWE classification in the annotation scheme of Vincze et al. (2011).

MWE and NE annotation was then conducted and inter-annotator agreement was measured, yielding a good k-score of 0.69. After analysing the annotation errors, the researchers found that most of them were due to conceptual differences and a lack of attention by the annotators. Nevertheless, using the MWE annotated corpus as training data had a positive impact on the performance of MWE recognition tasks. The MWE annotation scheme and the type of NLP applications in which MWE LRs were integrated can also be adopted with modifications to suit the objectives of the current research.

2.4.18 An automatic collocation extraction from a corpus (Saif and Aziz, 2011).

This study applied the automatic model in the extraction of multiple types of bi-gram AMWEs based on the evaluation of four AMs (Log-Likelihood Ratio, chi-square, Pointwise Mutual Information, and Enhanced Mutual Information). The corpus that was used consisted of a collection of newspapers texts compiled from various online resources and the extraction model includes two processing phases; candidate identification and ranking. The first stage involved generating the candidates and filtering using the n-gram model to extract bigram candidates based on the structural patterns of the AMWEs used, as shown in Figure 2.9.

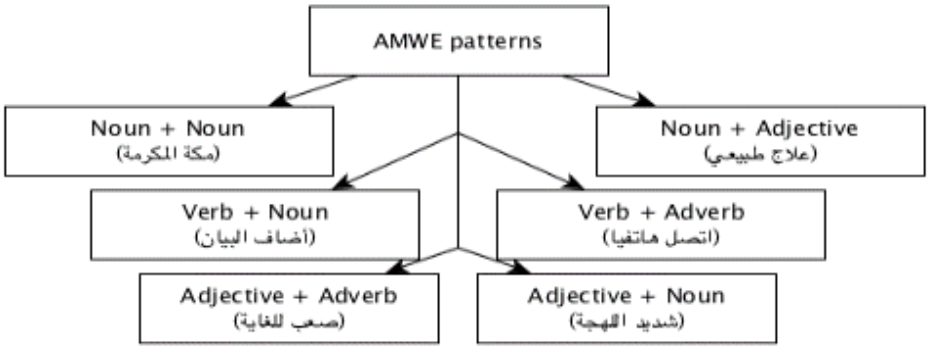


Figure 2.9: The structural patterns of AMWEs (Saif and Aziz,2011).

The extraction experiment first generates multiple lists of AMWEs that represent the selected morphosyntactic patterns with a total of more than 5k candidates. The filtering task then involves the removal of irrelevant candidates based on linguistic and statistical criteria. The ranking phase reorders the generated list of bigrams in descending order according to the selected AMs. The evaluation was based on the n-best evaluation method by measuring the best performing AM in ranking true AMWE candidates (Evert, 2004). The evaluation results show that Log-Likelihood Ratio and Enhanced Mutual Information achieve the best performance scores when ranking the n-best list of bigrams. In the thesis, the intention is to implement a hybrid model in MWE discovery which includes the use of AMs in extracting AMWEs; it will therefore be useful to compare the results with the findings reported in this study and explore whether the performance of these AMs might change with different morphosyntactic patterns and experimental settings for extraction.

2.4.19 An Arabic multiword expressions repository (Hawwari et al., 2012).

Hawwari et al. (2012) constructed a list of AMWEs comprising a collection of multiple existing AMWE dictionaries (Abou Saad, 1987; Seeny et al., 1996; Dawood, 2003; Fayed, 2007). The final list consists of 4,209 MWEs which were automatically tagged with the parts of speech tagger MADA (Nizar and Habash, 2010b; Roth et al., 2008). The MWEs were manually organised into several classifications according to their syntactic constructions. The N-N and V-N constructs constitute the dominant part of the extracted list with more than 3k items. Figure 2.10 presents the number of AMWEs in multiple morphosyntactic patterns.

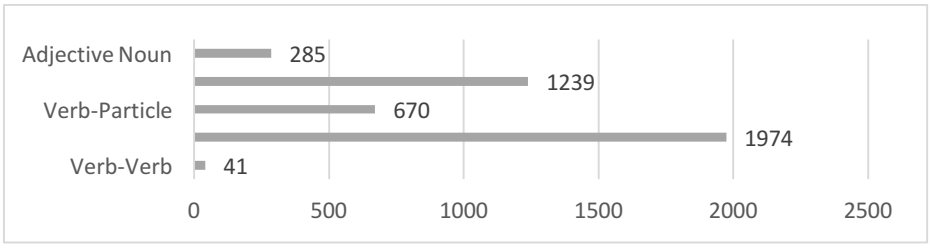


Figure 2.10: The distribution of AMWEs based on their construction classes (Hawwari et al., 2012, p. 26).

The primary goal of this list was to learn how to categorise new MWEs in large corpora statistically. The researchers developed a pattern-matching algorithm for detecting MWEs 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. Table 2.21 presents the results of the MWE annotation of the AGW corpus. The manual evaluation of a sample from the MWE annotation reveals an encouraging result with a high degree of accuracy.

Table 2.21: Annotated AMWEs by class.

MWE Construction	Number
Verb-Verb	576
Verb-Noun	64,504
Verb-Particle	75,844
Noun-Noun	316,393
Adjective-Noun	23,814

The developed list of AMWEs was then used in another study by Bar et al. (014b) to improve the performance of machine learning-based automatic identification and classification of AMWEs in the running text.

2.4.20 The lexical mark-up framework (Francopoulo, 2013)

Motivated by the efforts of the International Organisation for Standardisation (ISO), a group of 60 researchers (LMF team) spent more than five years constructing multilingual standards for representing LRs for NLP and Machine Readable Dictionaries (MRDs), which became known as LMF.

Standardisation plays a principal role in the reusability, development, distribution and evaluation of LR. Thus, the ultimate objective of LMF is to establish a

representational model that includes the minimum components that might constitute a consensus in the multilingual lexical representations of computational LRs. Odijk (2013) describes the core model for representing LRs designed by Unified Modelling Language (UML). Figure 2.11 shows the main representational classes included in the core LMF model. Each class comprises several attributes containing essential information related to the LR. For instance, the global information class consists of administrative information such as language coding. Furthermore, the LMF has several extension packages which can be used when needed to represent various linguistic descriptions. These are as follows:

Morphology.

Machine-Readable Dictionary.

NLP syntax.

NLP semantics.

Multilingual notation.

NLP morphological pattern.

NLP multiword expression pattern.

Constraint expression.

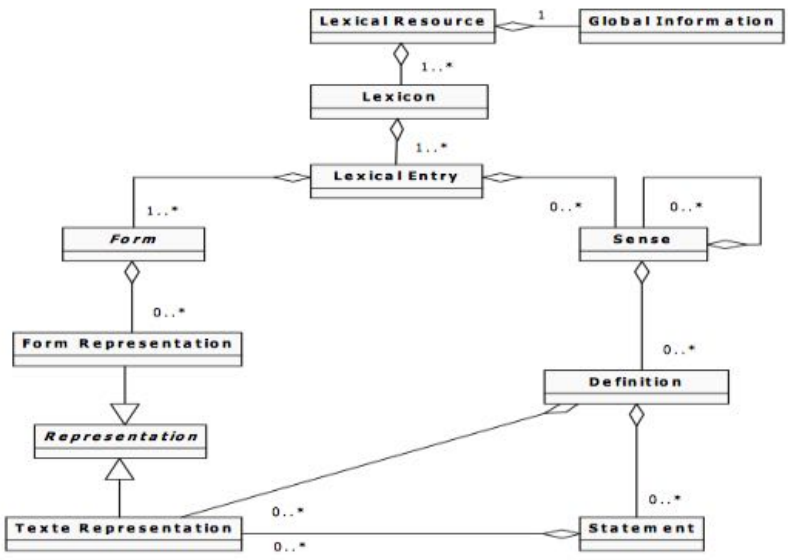


Figure 2.11: The core model of LMF (Francopoulo, 2013, p. 21).

The most relevant part of LMF is the MWE extension which can be used as part of the representational model in the current research. Figure 2.12 shows the NLP multiword expression pattern extension of LMF.

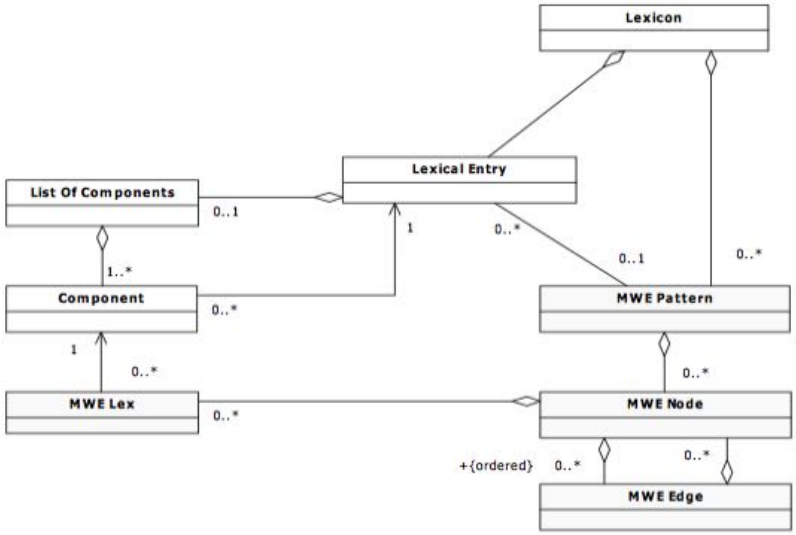


Figure 2.12: The MWE pattern extension for LMF (Francopoulo, 2013, p. 37).

Regarding the implementation for LMF in SA LRs, the work of Khemakhem et al. (2013) provides an example of previous efforts in this area in which they modelled the distinctive fundamental properties of SA lexicons in the LMF standards.

2.4.21 Lexical Semantic Analysis in Natural Language Text (Schneider, 2014).

This study developed a comprehensive English MWE annotation scheme to identify multiple types of MWE constructions; this was applied to a 56,000-word corpus of English. The annotation focused on three main properties of MWEs which are as follows:

heterogeneity—the annotated MWEs are not restricted by syntactic construction;

shallow but gappy grouping—MWEs are simple groupings of tokens, which need not be contiguous in the sentence;

expression strength—the most idiomatic MWEs are distinguished from (and can belong to) weaker collocations (Schneider, 2014b, pp. 46–47).

The MWE annotation results produced 460 morphosyntactic patterns of English MWEs, 73% of which consisted of two tokens. Table 2.22 presents the most common

MWE patterns. Schneider also found 2,378 types of MWE which reflects their heterogeneous nature.

Table 2.22: The most frequent patterns of English MWEs (Schneider, 2014).

MWE pattern	example
common noun–common noun	customer service
proper noun–proper noun	Persian_deity
verb-preposition	work with
verb-particle	look_for
verb-noun	take time:
adjective-noun	family~owned company
verb-adverb	call back

The most frequent instances of MWEs were "highly recommend(ed)", "customer service", "a lot", "work with", and "thank you", while the longest MWE consisted of 8 lexemes, such as "don't get caught up in the hype" and "don't judge a book by its cover". The annotation scheme and its implementation on English provides a valuable resource and methods for exploring the various linguistic properties of MWE knowledge. In the current research, one of the primary objectives is to propose an intensive representational model that describes the diverse linguistic characterisation of AMWEs in SA.

2.4.22 Classification and Annotation of Multiword Expressions in Dialectal Arabic (Hawwari et al., 2014).

Hawwari et al. (2014) developed a framework for classifying and annotating Egyptian AMWEs. Their research sought to build an intensive lexical resource for dialectal Egyptian AMWEs, enriched with comprehensive linguistic annotations that include phonological, orthographic, semantic, morphological, syntactic, and pragmatic information (ibid, p.49). The list is composed of 7,331 MWEs compiled from corpora and AMWE dictionaries. The annotation scheme described in Figure 2.13 contains 11 main linguistic features of MWEs, and a set of these features is also divided into sub-classifications. For instance, the semantic field class includes seven categories (e.g., social relations, occasions, and occultism).

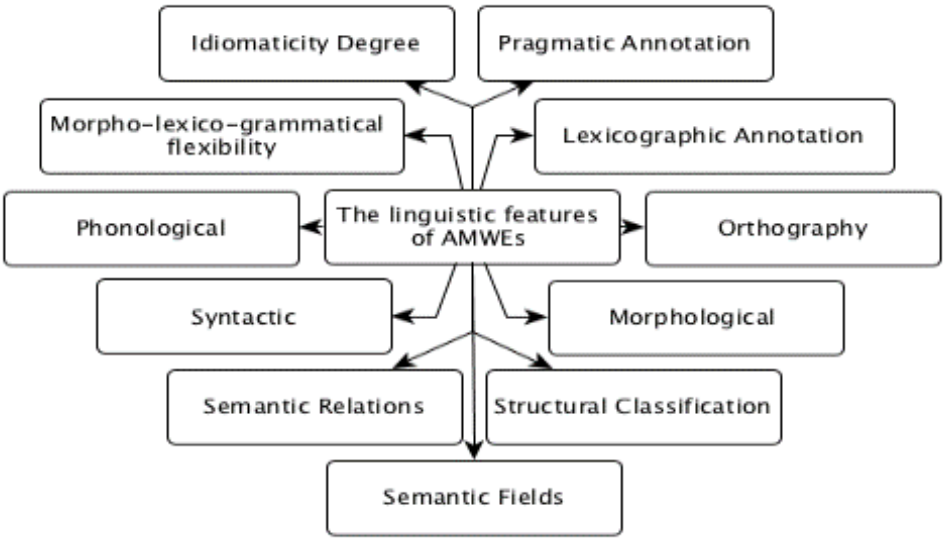


Figure 2.13: The Linguistic Features of Egyptian MWE annotation.

The developed framework built on previous research that has been applied to other languages such as Japanese (Shudo et al., 2011). Calzolari et al. (2002), then attempted to establish best practice and recommendations for representing MWEs in computational LRs.

2.4.23 A lexicon of multiword expressions for NLP (Tanabe et al., 2014).

This work was devoted to constructing an extensive lexicon of idiosyncratic Japanese MWEs; the dictionary contained 111k items which were extended with the MWE variants to 820k expressions. The comprehensive LR took decades to finish and was primarily divided into two main sub-lexicons: function and content MWEs. The first includes multiple types of MWEs such as phrasal verbs, light verb constructions, compound verbs, and compound nouns. The second comprises content MWEs such as discourse-relation-markers, complex sentence-connectives, and complex sentence-adverbs (e.g., in English, in other words, however, interestingly).

In defining MWE, the study focused on two main criteria, semantic non-compositionality and the strong statistical association between component words. The length of MWEs ranges from 2 to 18 lexemes with most expressions ranging from 2 to 4 component words. The extraction methods were based on manual collection from

various LRs such as newspapers, dictionaries, and journals. The LR consists primarily of 4 main components, which are as follows:

A large notational, syntactic, and semantic diversity of contained expressions.

A detailed description of the syntactic function and syntactic structure of each entry expression.

An indication of the syntactic flexibility of entry expressions (i.e., the possibility of additional, internal modification of constituent words).

An all-in-one architecture with uniform encoding schemas for each MWE (ibid, p. 1318).

The representation system for MWEs consists of seven main categories that contain various linguistic features of the phrase, as shown in Figure 2.14.

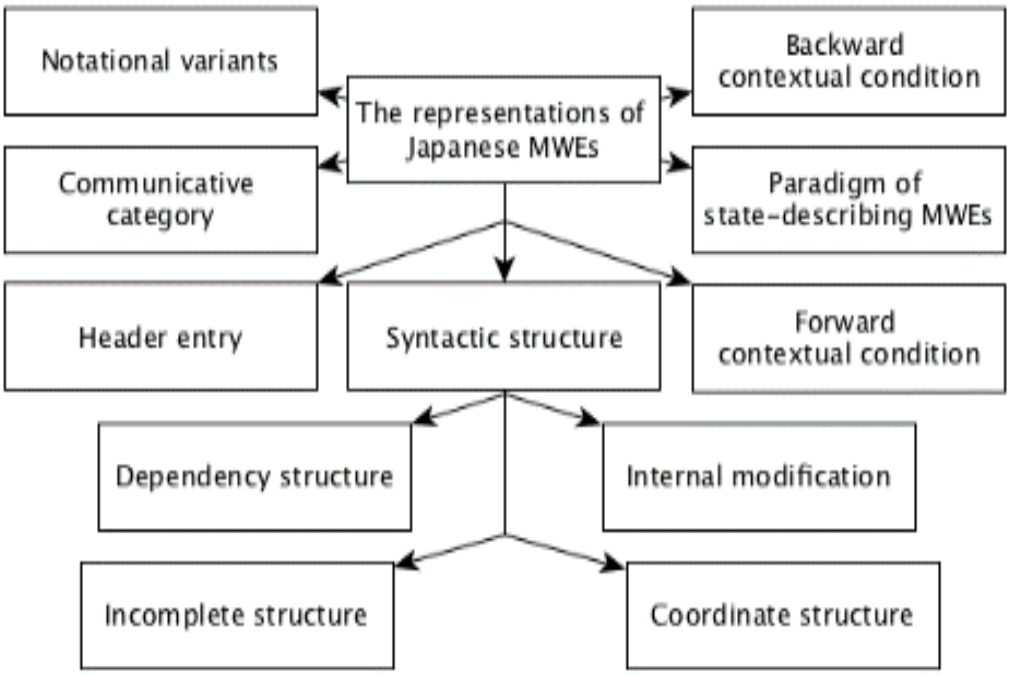


Figure 2.14: The representation model of the Japanese MWE lexicon.

However, each class of these linguistic features involves several other sub-classifications, as in the four main types of syntactic annotation. For instance, the first category is devoted to representing the syntactic constituents of the expressions based on the dependency grammar framework, as shown in Figure 2.15. This provides a syntactic dependency tree for the Japanese phrase that translates into "what will be, will be.".

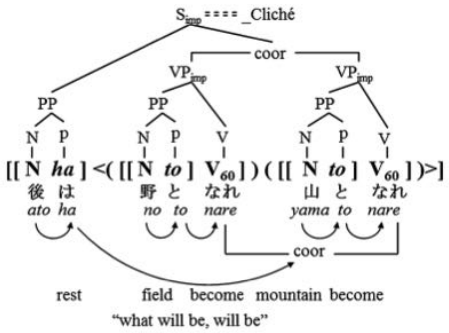


Figure 2.15: Syntactic representations of the Japanese structure "what will be, will be.".

Because the intention in the current research is to build a comprehensive representation model for AMWEs, this study along with others should be taken into consideration; it is also important to investigate whether the linguistic features adopted for Japanese and other languages could be used to represent AMWEs.

2.4.24 A repository of variation patterns for Arabic modal multiword expressions (Al-Sabbagh et al., 2014)

This study presents a domain specific type of research on AMWEs that proposes an annotation scheme for modality meanings and subcategories for clauses and verbal phrases in SA. The scheme was applied to a corpus of 1,704 raw tweets. In the annotation process, the project faced several challenges related to the distinctive properties of SA which were as follows:

the complexity of the Arabic modality paradigm

the lexical and semantic ambiguity of Arabic modality triggers

implicit scopes

word order flexibility

potential long dependencies between triggers and their scopes (Al-Sabbagh et al., 2014, p. 412).

The annotation was implemented using semi-automatic methods, which first involved automatic identification based on dictionary matching. The annotators then marked each modality trigger in the corpus along with its meaning, scope type, and span. The study used an Arabic modality lexicon which was collected manually by the authors;

this represented various meanings such as epistemic, sensory, reported, and abilities. Although this thesis has a different perspective, the annotation scheme and procedures applied by Al-Sabbagh et al. (2014) may be valuable in investigating the linguistic behaviour of modality expressions included in the proposed lexicon of AMWEs.

2.4.25 Extraction of Time-sensitive Arabic multiword expressions from social networks (Daoud et al., 2016).

This work presents another example of a domain-specific study on time-sensitive AMWEs. It involved using a statistical model to extract AMWE from a corpus that contains more than 15 million tweets. In the extraction experiment, bigram and trigram sequences were retrieved using the statistical model along with a search for other keywords and regular expressions. Figure 2.16 presents an overview of the experimental procedures implemented in this study.

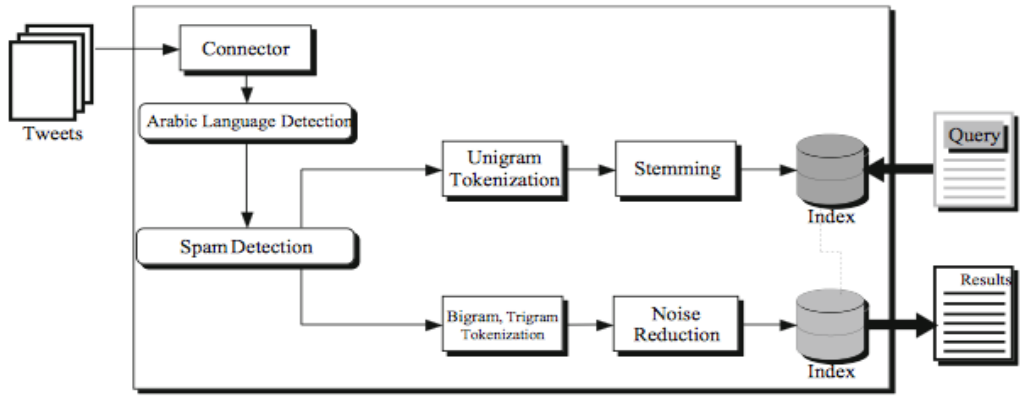


Figure 2.16: An overview of the experimental procedures adopted by Daoud et al. (2016, p. 255).

As shown, the extraction steps included the following. First, using the Twitter API, the researchers extracted tweets for 25 days and then the non-Arabic and spam tweets were removed based on specific criteria. Second, the corpus was tokenised and the bigram and trigram sequences were extracted, following which two indexes were created using the Lucene toolkit[22] The first employed an Arabic stemming analysis to search for potential AMWE items and the second included all the extracted AMWE

[22] Lucene is an open source text search engine; for more details see: https://lucene.apache.org/.

candidates. Evaluation of the AMWE extractor was based on measuring the quality of a sample from the extracted candidates representing various frequency levels. The evaluation result shows that the extractor works best with high frequency items with a 92.6 % precision score. Table 2.23 presents several examples of the most common AMWEs found in this study.

Table 2.23: Examples of the most frequent AMWEs (Daoud et al.,2016).

AMWE examples		Translation
عاصفة الحزم	ʿāṣifat alḥazm	the Storm of firmness
وزارة التعليم	wizārat attaʿlīm	Ministry of Education
الدولة الإسلامية	addawla alʾislāmiyya	the Islamic state
الجيش الصفوي	aljayš aṣṣafawī	the Safavid army

In the current research, multiple types of AMWEs will be covered to reflect the heterogeneity of this phenomenon in SA. Moreover, the extraction of discontinuous AMWEs which constitute an essential part of AMWE knowledge will also be taken into consideration.

2.5 Summary

In this chapter the most relevant works to the thesis have been briefly reviewed, beginning with a survey of the most common MWE extracting methods based on linguistic, statistical or hybrid models that have used a variety of manual and automatic discovery techniques from raw and annotated large corpora. In addition, a list was presented of the most relevant previous works on MWE LR lexicons and representations implemented in SA or other languages.

Piao et al. (2005, p. 378) emphasise that ‘Indeed, although numerous knowledge-based symbolic approaches and statistically driven algorithms have been proposed, efficient MWE extraction remains an unsolved issue'. Based on the works reviewed in this chapter this statement remains valid, especially in the context of AMWEs which have many linguistic features that pose serious challenges for computational processing.

Thus, this thesis will contribute to remedying this deficiency by implementing several MWE extraction models to build an intensive AMWE lexicon that can be used in several NLP applications. The current research adopts the hybrid approach to AMWE

extraction, which utilises statistical and linguistic models based on well-established quantitative and qualitative criteria.

To the best of the researcher's knowledge, no comprehensive computational lexicon of AMWEs has been attempted for various NLP applications. Hence, the study seeks to fill this knowledge gap by developing a corpus-driven lexicon of AMWEs that reflects the heterogeneous nature of this linguistic phenomenon in SA.

Furthermore, a comprehensive framework and representational model for AMWEs will be constructed that describes the distinct linguistic properties of AMWEs so that the declarative knowledge of MWE can be converted to imperative descriptions that are beneficial for multiple NLP tasks. In general, previous studies of AMWEs have presented a general description of approaches to AMWE extraction and have provided an explanation for the linguistic specifications of AMWEs that will be both beneficial and crucial for the current research.

3 Conceptual Framework for Arabic Multiword Expressions

3.1 Introduction

This chapter is devoted to addressing the first research question regarding the type and concept of targeted AMWEs and their distinctive linguistic characteristics. It begins by providing a general background to SA as this variant of Arabic was selected as the targeted genre[23] in the current work. This will be followed by a description of the adopted definitions and terminology utilised to place the thesis in a specific scope and context. The relevant linguistic or computational terms will be described and a framework presented that illustrates in-depth the linguistic characteristics of targeted AMWEs at various levels of analysis. Finally, this chapter concludes with a review of the existing typologies of MWEs and describes the AMWE taxonomy adopted in this thesis.

3.2 General Background on Standard Arabic

ANLP research is both a stimulating and challenging area because Arabic has a complicated linguistic system and a rich and ancient cultural and literary heritage. Arabic is believed to be the fourth most commonly spoken language worldwide with more than 395 million[24] native speakers. It is also the religious language of Islam and

[23] Arabic has been a living language for more than two thousand years and the spread of Arabic speakers throughout the world as well as the influence of other languages has led to a wide range of variation in uses of Arabic which, in most cases, are considered dialects of SA. However, in extreme cases, Arabic variants are considered an entirely different language by linguists as is the case in the Siculo-Arabic or Maltese. Egyptian, Western, Iraqi and Gulf are examples of colloquial Arabic dialects used mainly in everyday speech and by most of the Arabic users of popular social media networks (e.g., Twitter, Facebook).

[24] Statistics on languages can be obtained from the Statista website: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide.

is used daily by more than 1.6 billion Muslims around the world. The estimated number of internet and social media Arabic users is approximately 184 million. It is also the official and first language in 26 countries, and its script, which was initially used in the 4th century, is used officially by 15 other modern languages (e.g., Persian, Pashto, Urdu).

Furthermore, in Arabic, it is not extraordinary for natives with essential literacy skills to read and interpret a book that was originally authored more than fifteen centuries ago because, as Farghaly and Shaalan (2009) state, 'at the historical level, CA has remained unchanged, intelligible, and functional for more than fifteen centuries' (p. 14).

In the context of this thesis, the term SA or Modern Standard Arabic refers to a specific variant of Arabic that is used primarily in written and formal spoken discourse. Habash (2010a) states that SA 'has a special status as the formal written standard of the media, culture and education across the Arab World' (p. 1).

A detailed linguistic description of SA is beyond the scope of this thesis, thus the aim is to shed light on the core properties of SA that make it distinctive from other languages and English in particular. The focus will also be on the linguistic features that pose challenges for several ANLP tasks at various levels of processing. This introduction paves the way for an explanation of the linguistic characteristics of AMWEs presented in sections 3.3 and 3.4.

For an intensive introduction to SA, the reader can consult a variety of comprehensive resources on Arabic linguistics (e.g., Darwish, 2014; Dickins and Watson, 1999; Abdou, 2011; Nizar Y. Habash, 2010a; Holes, 2004; Ryding, 2005; Badawi et al., 2013; Farghaly and Shaalan, 2009; Rosenhouse and Versteegh, 2006; Fehri, 2012) .

3.2.1 Distinctive properties of standard Arabic

The following subsections present brief descriptions of the core linguistic features of SA that pose various challenges in most ANLP tasks.

3.2.1.1 Arabic script

Arabic script has several key features that should be considered in automatic processing tasks. The first of these is the direction of writing which in Arabic runs

from right to left. This poses a problem when integrating Arabic into NLP tools that do not consider this in the software construction process.

The absence of capitalisation also poses challenges, especially in the computational processing of named entities. Furthermore, there are no strict rules of punctuation which make various processing tasks related to the identification of sentence boundaries much harder. In most Latin script-based languages the uppercase feature assists greatly in the improvement of many NLP tasks, such as the identification of NEs in the running text, but in Arabic there is no such feature.

Another feature is related to the representation of short vowels by diacritics, which are a set of marks above or under the letter. These play a vital role in selecting the correct vocalised form of the word; ignoring this feature therefore eliminates the precision of system outputs. Table 3.1 shows the diacritic marks and illustrates their significant effect by presenting several vocalised forms of the word علم *ʿalam* where these marks are the only means of differentiation, especially when out of context.

Table 3.1: Arabic diacritic marks with examples of vocalised variations of words.

Diacritic marks	◌َ	◌ِ	◌ُ	◌ْ	◌ّ
Arabic words	عَلَم	عِلْم	عُلِمَ	عَلَّمَ	عَلِمَ
POS	Noun	Noun	Verb	Verb	Verb
Transliteration	ʿalam	ʿilm	ʿulima	ʿallama	ʿalima
Translation	Flag	Science	Known (passive voice)	Taught (past)	Knew (past)

Arabic is also generally considered a phonetic script, which means that each letter uses one-to-one mapping with its counterpart sound. Moreover, Arabic does not require a combination of two letters or more to represent a single sound. It is important to note that several letters in Arabic have the same basic shape and merely add the dots as distinguishing marks between them, as can be seen in this set of three letters (خ – ح – ج) (ث – ت – ب) (ba – ta – ṯa) (ja – ḥa – ḫa).

A further distinctive property relates to the variant forms of letters written in Arabic, in that most letters have multiple written shapes according to their position in the word. Arabic contains 28 letters, each of which has at least three different written forms. Table 3.2 presents examples of letters of different shapes. Furthermore, several

letters have more than three shapes such as the Hamza or Alif letter which has six various written forms (آ – ء – ئ – ؤ – أ – إ) (ʾia – ʾa – ʾa – ʾa – aa – aʾaa).

Table 3.2: Different shapes of Arabic letters based on their position.

Position		Isolated	Initial	Medial	Final
Example Letters					
كاف	kāf	ك	كـ	ـكـ	ـك
عين	ʿayn	ع	عـ	ـعـ	ـع
سين	siyn	س	سـ	ـسـ	ـس
هاء	hāʾ	ه	هـ	ـهـ	ـه

The final point to make concerns the normalisation of Arabic script, which is an essential pre-processing step in most ANLP tasks. The main reason for normalisation is to cover the variations in Arabic script, especially when processing letters with various forms such as Alif. According to Habash (2010:22), the normalisation of Arabic script usually includes the following subtasks:

Tatweel removal: The Tatweel symbol (ـ) is removed from the text[25] .

Diacritic removal: Because diacritics occur so infrequently, they are considered noise by most researchers and are simply removed from the text.

Letter normalisation: Four letters in Arabic are misspelt so often when using variants that researchers find it more helpful to make these variants entirely ambiguous (normalised). The following are the four letters in order, from the most commonly normalised to the least commonly normalised (the first two refer to what most researchers do by default, the last two are less commonly applied).

The Hamzated forms of Alif (آ - إ - أ) are normalised to bare Alif (ا).

The Alif-Maqsura (ى) is normalised to (ي).

The Ta-Marbuta(ة) is normalised to (ه) .

The non-Alif forms of Hamza (ئ - ؤ) are normalised to the Hamza letter(ء)'.

[25] This symbol is used in Arabic script for decorative purposes, as can be seen in these two words before and after the removal of Tatweel :عمـــار - عمار.

3.2.1.2 Non-concatenative morphology

Computational morphological processing of Arabic lies at the heart of most ANLP research. This is because one of the primary distinguishing characteristics of Arabic is a rich and complex derivational and inflectional morphology which poses various open research problems in most NLP tasks.[26] The morpheme is defined as a minimal grammatical component. Unlike English Arabic has non-concatenative morphology (McCarthy and Prince, 1994). McCarthy (1981, p. 375) describes this as follows:

> *It has long been known that at its basis there are roots of three or four consonants which cluster around a single semantic field, like ktb 'write'. Specific changes in these roots, like gemination of the middle radical in (lb), yield derivatives such as causative or agentive. Moreover, some vowel patterns seem to bear consistent meaning, like the difference in stem vocalism between active kataba and passive kuitiba.*

Habash and Rambow (2005b, p. 573) also explain the complexity of the Arabic morphological system and how it differs in its entirety from English morphology, stating that:

> *Arabic is a morphologically complex language. The morphological analysis of a word consists of determining the values of a large number of (orthogonal) features, such as basic part-of-speech (i.e., noun, verb, and so on), voice, gender, number, information about the clitics, and so on. For Arabic, this gives us about 333,000 theoretically possible specified morphological analyses, i.e., morphological tags, of which about 2,200 are used in the first 280,000 words of the Penn Arabic Treebank (ATB). In contrast, English morphological tagsets usually have about 50 tags, which cover all morphological variation.*

Thus, Arabic is primarily a root-driven language, and Arabic morphemes may have boundaries within this. Morphological analysis should therefore consider three levels of analysis for the Arabic word; the root, the vocalism, and catenative affixation

[26] The related computational approach to Arabic morphology will be illustrated where appropriate in chapters 4, 5 and 6 of this thesis.

(McCarthy, 1981; Farghaly, 1987). This feature is a dominant phenomenon in Arabic, where several words within a specific semantic field belong to one consonantal discontinuous root radicals, as shown in Table 3.3 which presents a list of examples of words related to the root (ك- ت- ب).

Table 3.3: List of Arabic words derived from the root K T B (ك - ت - ب)[27].

katab	كَتَب	kitāb	كِتَاب
kitāba	كِتَابَة	kutub	كُتُب
kitba	كِتْبَة	muktatibūn	مُكتَتِبُون
katb	كَتْب	kutayyib	كُتَيِّب
maktūb	مَكتُوب	kutubī	كُتُبِيّ
kātab	كَاتَب	kuttāb	كُتَّاب
mukātaba	مُكَاتَبَة	katātīb	كَتَاتِيب
ʾaktab	أَكتَب	kitābāt	كِتَابَات
mutakātib	مُتَكَاتِب	katāʾibiyyah	كَتَائِبِيَّة
istaktab	اِستَكتَب	maktabī	مَكتَبِيّ
istiktāb	اِستِكتَاب	maktaba	مَكتَبَة
mustaktib	مُستَكتِب	maktabāt	مَكتَبَات

Attia (2008, pp. 31–33) lists the sources of genuine morphological ambiguities in SA as follows:

Orthographic alternation operations (such as deletion and assimilation) frequently produce inflected forms that can belong to two or more different lemmas.

Some lemmas are different only in that one has a doubled sound which is not normally made explicit in written form.

Many inflectional operations involve a slight change in pronunciation without any explicit orthographical effect due to a lack of short vowels (diacritics).

Some prefixes and suffixes can be homographic with each other.

[27] These examples were extracted from the ElixirFM Functional Arabic Morphology System (Smrz and Bielický, 2010).

Prefixes and suffixes can accidentally produce a form that is homographic with another full form word.

There are also the usual homographs of uninflected words with/without the same pronunciation; these have different meanings and usually different POSs.'

Finally, the dilemma of word classes in Arabic, which is a controversial topic in the literature, begins with a minimum of three basic classes (Noun-Verb-Particle), which are the dominant cases in CA linguistic literature, to more than 2000 possible POS tags, as is the case in the Penn Arabic Treebank (ATB) (Maamouri and Bies, 2004). Another comprehensive POS (SALMA system) developed by Sawalha (2011) consists of 22 main features, each of which includes several subcategories for various morphological representations. Details about the adopted POS for SA in the thesis will be described in research experiments when relevant. Table 3.4 presents examples of the POS tagset developed by Attia et al. (2017 pp. 8–13) based on the main POS used in universal dependency grammar representations.

Table 3.4: Arabic POS tagset with examples based on POS of universal dependency.[28]

POS	Examples	
ADJ: adjective	مجتهد	mujtahid
ADV: adverb	أيضا	ʾayḍan
ADP: preposition or subordinating conjunction	من ، أمام	min , ʾamām
CONJ: coordinating conjunction	لكن	lakin
DET: determiner	بعض	baʿḍ
INTJ: interjection	كلا	kalā
NOUN: noun	كتاب	kitāb
NUM: numeral	عشرون	ʿašrūn
PART: particle	هل	hal
PRON: pronoun	أنا	ʾanā
PROPN: proper noun	أحمد	ʾaḥmad
PUNCT: punctuation	"" ،.	
SYM: symbol	$ # @	
VERB: verb	سمع	samʿ

3.2.1.3 An agglutinative and pro-drop language

Agglutinative and pro-drop are two core linguistic features of SA that will now be briefly discussed. Arabic agglutination means that the word or token structure is complicated because it largely consists of a mix of affixes and clitics. This plays a crucial role in understanding Arabic at various levels (e.g., words, phrases, sentences). Clitics and affixes represent different functions and can be analysed as POS with a syntactic function or in other cases as markers for tense, gender, person, number, and voice. Thus, in Arabic, it is difficult to consider the white space as the word boundary. Nonetheless, what appears initially as one word can be analysed as a full sentence in SA. For example, Figure 3.1 presents a morphological analysis of the one token sentence فَأَسْقَيْنَاكُمُوهُ *faʾasqaynākumūhu* 'We gave it to you to drink' which is decomposed into five morphemes (ignoring the internal morphological structuring of

[28] A full reference for universal dependencies can be accessed at:
`http://universaldependencies.github.io/docs/u/feat/all.html`

asqay) with a specific syntactic function for each. Therefore, it is hard to define a word in Arabic as a string of characters delineated by spaces. This complex system of affixes and clitics in Arabic makes tokenisation and POS tagging a very challenging task and this has a considerable effect on most NLP processing tasks, especially MWE discovery, which will be illustrated in-depth later in the thesis when relevant.

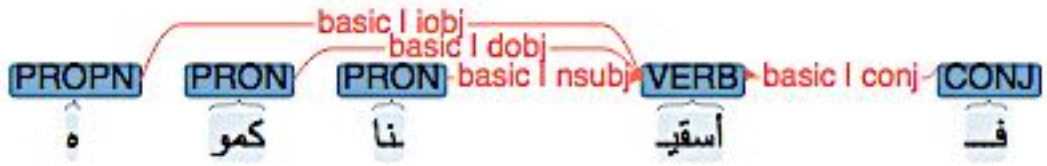


Figure 3.1: A morphological and syntactic analysis of the sentence *faʾasqaynākumūhu.*

Another core feature of SA is the pro-drop property which can also be found in many modern languages.[29] Arabic permits the dropping of the subject pronoun and the contraction of a sentence without a subject which causes different types of syntactic ambiguity. An example of this can be seen in the second part of sequences of Verb-Noun phrases where there is no pronoun subject in the second phrase (null-subject), yet the meaning is still preserved by the sentence context. As we can see in the equivalent English sentence, the subject 'he' remains in the second phrase and cannot be removed. An intensive discussion and examples of this feature in Arabic can be found in Fehri (2012), Altamimi (2015), and Alnajadat (2017).

ساعد صديقك، يساعدك[30]

sāʿid ṣadīqaka, yusāʿiduka

Help your friend, so he helps you

[29] The pro-drop or pronoun drop language is defined as 'A language in which an empty subject position that has been motivated by the projection principle and which has pronominal, i.e. referential, properties can appear in a finite sentence. Examples of such languages are Italian and Spanish, but not English, German, or French. For example, compare Italian [pro mangia] with English *[pro eats] for 'he eats'. The pronoun "he" cannot be dropped in English'(Bussmann, 2006, p. 948).

[30] Traditional CA grammarians express this phenomenon by assuming a hidden pronoun (الضمير المستتر) which refers to the removed subject.

A final issue relates to the affixes and clitics system in SA. In the literature, Arabic clitics are optional and do not change the core meaning of the attached string; they are also divided into proclitics and enclitics which are used to denote several functions in the discourse. Table 3.5 presents examples of the frequent clitics used in SA.

Table 3.5: Examples of Arabic clitics and their functions.

Clitics	Class	Function	Example	English
أ	Particle	interrogative	أسمعت ʾasmaʿtu	yes/no question
و	Conjunction	Coordination Connection accompaniment	قلم وكتاب qalamun wakitāb	and and with
ك	Particle	preposition	كالقمر kalqamar	such as, like
س	Particle	Future preposition	سننجح sananjaḥ	will
ال	Determiner	definite article	السلام assalām	the

However, affixes are obligatory and represent inflectional morphology in SA. They are attached to various word classes to indicate they are inflected for aspect, mood, voice, person, gender, and number. In SA, a person has three values: speaker, addressee and other or third person, while gender has two values: masculine or feminine, and number has three values: singular, dual, or plural. Table 3.6 presents examples of common affixations used in various inflectional forms.

Table 3.6: Examples of SA affixations.

Affix	Class	Functions	Word Example	
			Medial	Final
ه	pronoun	third person	أعلنها ʾuʿlinuhā.	أنه ʾannahu
ت	pronoun	addressee	رأيتهم raʾaytuhum	رأيت raʾayt
ي	pronoun	feminine addressee	تساعدين tusāʿidīna	ساعدي sāʿidī
ا	pronoun	dual	يساعدان yusāʿidāni	ساعدا sāʿadā
ون	sound plural	masculine plural		مجتهدون mujtahidūn
ات	sound plural	feminine plural		مجتهدات mujtahidāt
ان	sound dual	dual		مجتهدان mujtahidān

In this section, the nature of SA morphology will be described only briefly as this topic is relatively complicated and interacts profoundly with other levels of linguistic

analysis in Arabic. Hence, a complete discussion of these issues is beyond the scope of the thesis.

3.2.1.4 Syntactic structure

In most morphologically rich languages the interaction between syntax and morphology is both substantial and complicated. This is because syntactic relations in SA are not restricted to merely exhibiting various ways of ordering words as they use internal morphological variations of words to express several syntactic phenomena. The syntax also interacts heavily with phonology because of the vocalised form of words - mainly when short vowels are added to the end of words - to indicate their grammatical cases. Hence, these morphosyntactic interactions result in one of the main features of SA sentence structure which is relatively free word order. This means that no strict or fixed order is required when making correct grammatical structures. However, in several situations, case ending markings play a significant role in selecting the meaning of a specific structure. Table 3.7 shows examples of different word orders in SA.

Table 3.7: Example of various word orders in SA sentences.

Order	Sentence example		
Verb-Subject-Object	الرسالةَ	الطالبُ	كتبَ
	arrisālata	aṭṭālibu	kataba
	the letter	the student	wrote
Subject-Verb-Object	الرسالةَ	كتبَ	الطالبُ
	arrisālata	kataba	aṭṭālibu
	the letter	wrote	the student
Verb-Object-Subject	الطالبُ	الرسالةَ	كتبَ
	aṭṭālibu	arrisaālata	kataba
	the student	the letter	wrote

In SA, sentences can be classified into two core types, which are as follows:

1- Verbal sentence: a sentence with the main clause beginning with a verb.

2- Nominal sentence: a sentence with the main clause beginning with a noun.

Examples of such sentences are presented in Table 3.7 where the first and third examples are verbal sentences and the second is a nominal sentence.

SA uses the same case system in CA which primarily includes three cases which are indicated by adding short vowel marks to the ends of words in written SA,[31] as shown with examples in Table 3.8.

Table 3.8: SA Arabic cases with examples.

Case	Marks	Examples	Grammatical Function
Nominative	ُ	كتبَ الطالبُ الدرسَ	The subject of a verbal sentence.
		kataba aṭṭālibu addarsa	The subject and predicate of a nominal sentence.
		الطالبُ مجتهدٌ	
		attālibu mujtahidun	
Genitive	ِ	من الماءِ	The object of a preposition.
		min almā'i	The second term of a genitive structure.
		نورُ العلمِ	
		nūru al'ilmi	
Accusative	َ	رأيتُ القمرَ	The object of a transitive verb.
		ra'ytu alqamara	The circumstantial accusative.
		وصل مبكراً	
		wasal mubakkiran	

According to Attia (2008, p. 176), syntactic ambiguity in SA has the following main sources: pro-drop feature, the flexibility of word order, diacritic ambiguity, and multifunctional nouns. It is worth noting that in Arabic there is no equivalent in the present of the English copular verb, so to construct an Arabic sentence similar to the English 'I am a student' only a pronoun and noun are needed in SA أنا طالب *ʾanā ṭālib'* without the use of the (to be) verbs. The agreement system in SA also presents several challenges. Three types of agreement are briefly described. First, the noun and their various modifiers must agree in definiteness, number, gender, and case. Second, in the Verb-Subject-Object structures, the verb must agree with its subject in gender only and is always used in singular form regardless of the subject number values. Third, in the Subject-Verb-Object, the verb must agree with its subject in person, gender, and number. As mentioned previously, this section makes several brief points about the most distinctive properties of SA that will assist in elucidating the following sections regarding the adopted definition of AMWEs and their linguistic properties.

[31] In SA, these marks are usually not written but native educated speakers pronounce them as a short vowel at the end of words.

3.3 Core concepts and definitions

After defining what is meant by SA and its core linguistic characteristics, the following subsections introduce the adopted definitions of the core concepts used in this thesis to place the work in a specific scope and context. It will begin with a brief note on terminology issues before illustrating the concept of AMWE and several other related terms that may be of relevance to the thesis.

3.3.1 A brief note on terminology

Terminology is a very complicated issue in the MWE literature. Wray (2002a) refers to a plethora of terms as more than 50 have been used in references to various phenomena which, in many cases, are considered duplicates in that the researchers have described the same phenomenon in different terms. Figure 3.2 lists examples of terms used in the literature.

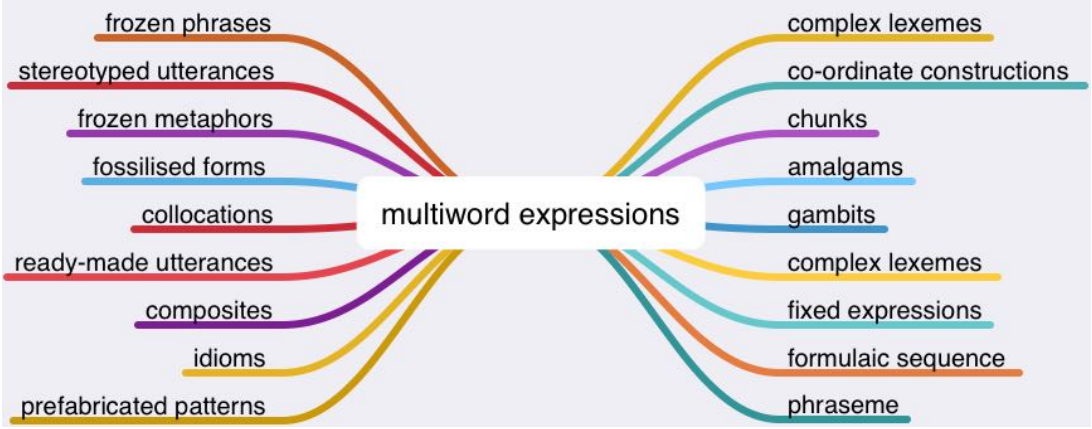


Figure 3.2: List of terms used in the literature to describe MWE phenomena.

Such a large number of terms is justified according to Granger and Paquot (2008) because 'the unwieldy terminology used to refer to the different types of multi-word units is a direct reflection of the wide range of theoretical frameworks and fields in which phraseological studies are conducted and can be seen as a sign of the vitality of the field'(Granger and Paquot, 2008, p. 13). However, to ensure consistency and coherency within this thesis, a specific term will be imposed and used throughout. Thus, the term MWE will be employed or AMWE when referring to Arabic expressions. It is assumed this term is the most relevant to the adopted definition of this phenomenon and is also the most widely used term in NLP literature. It can also be used as an umbrella to denote various types of MWEs in general. However, in rare

cases, other terms such as FSs, constructions, co-occurring words, and collocations might also be used interchangeably.

3.3.2 What are AMWEs?

When attempting to define MWEs, the heterogeneous nature of this phenomena in human languages at different linguistic levels becomes clear (e.g., morphology, syntax. and semantic). Hence, it is hard to find a consensus in the literature on what MWEs are as many definitions have been suggested (e.g., Sag et al., 2002; Wray, 2002; Baldwin, 2005; Durrant, 2008; Abdou, 2011; Ramisch, 2012; Constant et al., 2017)[32]. This is due to the complex linguistic properties of MWEs; like the well-known tale about blind men touching an elephant, every researcher attempts to demonstrate his or her understanding of these complicated related phenomena. For instance, in CL and NLP the term 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 multiword expressions is beyond the reach of this research, particularly for morphologically rich languages such as Arabic. In this thesis, a practical definition of AMWE will be employed which covers all types of expressions targeted in this research. The adopted definition is based on the research objectives which focus on the Arabic expressions that are most valuable in eliminating multiple types of language ambiguity problems in various NLP tasks that are caused mainly by inadequate MWE knowledge. The primary focus on the concept of MWE in terms of phrases might pose various challenges in traditional word by word computational processing. Notably, these types of expressions, which have been found in LP literature to be the most beneficial part of the formulaic language, substantially contribute to enhancing fluency, proficiency, and thus comprehension among second language learners.

For the purpose of this research, the working definition of multiword expressions adopted is mainly based on Baldwin and Kim's (2010) concept of MWE which in turn is based on Sag et al.'s (2002) definition of MWEs, which is as follows:

[32] See Appendix B for a list of frequently cited definitions of MWE in the literature.

> *Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. (Ibid., 2010, p. 269)*

This is the most appropriate definition for describing the multiple types of component lexemes targeted in this research. This definition includes several core features of MWEs that are mostly related to the various linguistic and statistical characteristics of this phenomenon. These will be illustrated in-depth with examples of AMWE in sections 3.4 and 3.5 of this chapter. Selecting the term lexemes[33]for this definition is essential because the question as to what constitutes the fuzzy notion 'word' in Arabic is a vexed one, and the word can usually be interpreted as the minimum element of vocabulary. Nonetheless, for SA, complete sentences can on many occasions be found in one space-delimited token as shown in examples in section 3.2.1.3.

3.3.3 Practical criteria for defining AMWEs

Based on the previously adopted definition of MWE and a comprehensive analysis of the targeted Arabic component lexemes in this project, this section presents a practical list of criteria for selecting different types of AMWEs that can be utilised in the computational and manual filtering component of AMWE extraction models. This section will also present excluded types of expression that are beyond the scope of the current research. The main AMWE criteria are as follows:

AMWEs consist of a minimum of two lexemes or more. In SA two or more lexemes MWEs can be found in one-string word which also consider as MWE in this research. Regarding the maximum number of lexemes in a MWE, there is theoretically no limitation in this study regarding MWE length; however, most of the AMWEs considered should not exceed six lexemes, as is the case in most MWE research.

[33] A lexeme is defined in linguistics as a 'Basic abstract unit of the lexicon on the level of language which may be realised in different grammatical forms such as the lexeme 'write' in 'writes', 'wrote', 'written'. A lexeme may also be a part of another lexeme, e.g. 'writer', 'ghostwriter', and so on' (Bussmann, 2006 p. 670).

Discontinuity: this means that AMWE can be continuous or discontinuous. In the experiments this criterion is accounted for by allowing the extraction of discontinuous phrases as illustrated later in section 3.4.2.

Idiomaticity and compositionality: most MWEs show a degree of idiomaticity that is apparent at multiple linguistic levels (e.g., lexical, syntactic, semantic, pragmatic). For instance, semantic idiomaticity denotes the semantic relation between the meaning of an MWE as a whole and its component parts when the meaning of an MWE cannot be explicitly derived from its constituted components and is called a non-compositional MWE.

Frequent recurrence or statistical idiomaticity. MWEs of various types tend to consist of commonly co-occurring words. Thus, markedly high frequency is a defined criterion for MWEs. This feature is also crucial because it is one of the most simple criteria to implement using computational discovery methods.

MWES are prefabricated units. Many definitions in the literature assert that MWEs are represented in our mental lexicon as a linguistic chunk rather than merely individual words (e.g., Isabelli, 2004; Schmitt, 2004; Wray, 2013). For instance, Wray (2002a) defines what she calls ‘FSs', as ‘a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar'. This criterion could be used when using native speakers’ intuitions when selecting or evaluating a list of extracted MWEs.

Arbitrariness or idiosyncraticity. In contrast to regularity, several MWEs might not conform to various language rules and show multiple types of linguistic arbitrariness This feature of MWEs is illustrated in the literature by the notion of institutionalisation (Wray, 2012; Garrao et al., 2008; Sag et al., 2002), which primarily refers to the emergence of MWE from intensive use of specific phrases to denote a particular function or notation. For instance, the phrase ‘good morning’ is considered a conventionalised indicator : “a polite greeting phrase to people in the morning time". The heavy use of this phrase in denoting this communicative function transforms it from normal status to institutionalisation status.

The components of MWEs in the proposed conceptual framework have certain types of syntactic relations; however, there are no constraints regarding the type of syntactic structures included in the study. Instead, all possible grammatical structures in AMWEs will be covered. Thus, syntactic analysis when available is vital in enriching the lexicon.

Hence, based on these criteria, most types of morphosyntactic constructions in MWE literature are included in the current research. These are as follows:

Nominal expressions.

Verbal expressions.

Adjectival expressions.

Adverbial, including prepositional, expressions

Other types of MWEs, namely proper nouns and MWTs, are excluded from the research because they are beyond the scope of this thesis. In the literature, a vast amount of research can be found that is exclusively devoted to covering these two linguistic phenomena, which are mostly referred to as named entity and terminology recognition, and extraction research areas. The list presented is not intended to be an exhaustive list of all MWE criteria, but a guide that includes the distinctive core properties of AMWEs and helps illustrate the adopted definition of AMWE.

In the current research, any MWE that meets at least one of these criteria is considered valid. This concept includes any semantically regular formulas that are not restricted to any syntactic construction or semantic domain. More details on the linguistic characteristics of AMWE with examples are presented in section 3.4.

3.3.4 Important related terms

The following subsections briefly provide a description of several terms that are frequently used in this thesis. However, a full discussion of these research fields is beyond the scope of this thesis.

3.3.4.1 Computational linguistics, natural language processing, and corpus linguistics

The core issues discussed in the thesis lie at the intersection of three research fields: CL, NLP, and corpus linguistics. In this section, the primary objectives of these fields will be clarified and the nature of the interactions between them explained. CL and NLP[34] are language engineering terms that overlap considerably in the literature, and the difference between them is fading to the extent they are being used interchangeably. The common object of both is the scientific study of natural languages which includes all levels of linguistic description and analysis from computational perspectives. However, several researchers prefer to use the term NLP, especially for applications-oriented research where the core focus is on building practical applications, algorithms, and software for NLP tasks. In contrast, CL refers to studies that have used computational methods for implementing linguistic-oriented solutions to various types of natural language problems. Nonetheless, the core objective of this field of research is 'to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech' (Jurafsky and Martin, 2007, p. 1).

Thus, NLP covers various areas of interest and in many cases interacts intensively with other related areas which fundamentally renders the nature of most research in this field interdisciplinary. Since the 1970s, several approaches have been used in NLP literature. Dale (2010) suggests these can be classified into four main research directions:

Classical symbolic approach.

Statistical or corpus-based methods.

---

[34] These two terms are used interchangeably in this thesis. Natural language also refers to 'languages which have developed historically and which are regionally and socially stratified, as opposed to artificial language systems, which are used for international communication or for formulating complex scientific statements. Natural languages differ from artificial languages particularly in their lexical and structural polysemy, the potential ambiguity of their expressions, and in their susceptibility to change through time' (Bussmann, 2006, p. 788).

NLP research based on machine learning, artificial neural network, or deep learning techniques.

A hybrid approach that incorporates the best practice of multiple approaches.

Given the recent and enormous number of NLP research studies utilising statistical and machine learning methods, Dale (2010) emphasises that 'these changes should not be taken as an indication that the earlier-established approaches are somehow less relevant; in fact, the reality is quite the opposite, as the incorporation of linguistic knowledge into statistical processing becomes more and more common'(pp. 3–4). The processing spectrum in a classical approach usually consists of a pipeline[35] of stages beginning from surface text tokenisation and ending with advanced in-depth semantic and pragmatic analysis.

Corpus linguistics is a large research field that overlaps with NLP in 'processing a wider range of discourse but at a restricted level of analysis (e.g. syntax or semantics)'(Rayson, 2002, p. 10). It can be defined as 'an area which focuses upon a set of procedures, or methods, for studying language' (McEnery and Hardie, 2011, p. 1). The availability of a significant amount of textual data and the computing power to process them in various ways enables researchers to develop new insights into languages and assists considerably in refuting or refining previous claims, theories, and hypothesises in several language-related disciplines. Since the development of the early Brown corpus in the 1960s (Leech, 1997), researchers have used different types of corpora in a wealth of research conducted in linguistics and social science-related areas.

In the literature, several classifications have been suggested for corpus linguistics. However, McEnery and Hardie (2011, p. 3) suggest there are six main features that can be used to distinguish different types of research in this area, which are as follows:

[35] 'Pipeline' here means the sequence order of processing, where the output of one stage is the input of the following stage. However, this is a point of controversy in the literature as many argue about the suitability of considering natural languages as separate parts that can be analysed sequentially. They argue that human languages should be viewed as a combination of phenomena that should be processed using paralleled nonlinear methods.

Mode of communication;

Corpus-based versus corpus-driven linguistics;

Data collection regime;

The use of annotated versus unannotated corpora;

Total accountability versus data selection;

Multilingual versus monolingual corpora'.

Because the AMWEs lexicon in this study is a corpus-based LR, several research methods and standards suggested by corpus linguistics will be used regarding corpora evaluation, annotation procedures, and various techniques for exploring language data, mainly related to extracting MWE and collections from raw and annotated corpora. Constructing a lexical resource, which is the primary aim of this research, can be placed at the intersections of these research fields because they provide the researcher with a wide range of resources and methodologies that ultimately assist in enhancing the AMWE LR. Huang et al. (2010 p. 15) emphasised that 'the importance of a multidisciplinary approach is recognised for lexical resources development and knowledge representation as acknowledged by many influential contributions to the field'. Hence, in this study the integration of several methods and techniques from multidisciplinary perspectives will be used to enhance the overall quality of the AMWE lexicon.

3.3.4.2 Language resources

Language resources (LR) in this thesis means any type of machine-readable language data and thus includes several forms of data that were constructed for various purposes in NLP or other language-related fields. These include various kinds of corpora, electronic lexicons, tree banks, morphological lexicons, and different types of MWE and phrase knowledge bases. It also includes ontologies which have recently been utilised in the development of various NLP semantic tools. Ontologies have much in common with the lexicon, as they include an inventory of concepts and terms associated through various types of relations, such as paradigmatic and syntagmatic relations. On the other hand, a lexicon, which is the type of LR constructed in this study can be defined as 'a collection of linguistically conventionalised concepts'

(Huang et al., 2010, p. 6). In another definition focusing on NLP-oriented lexicons, they as defined as 'digital knowledge bases that provide lexical information on words (including multi-word expressions) of a particular language' (Gurevych et al., 2016). The lexical entries usually include several types of linguistic metadata, as will be illustrated in the following section.

Ontologies and lexical resource knowledge can be linked and combined to enhance their coverage and potential applications. Several types of lexicon have been mentioned in the literature and these can be classified according to several linguistic features including monolingual, multilingual, single word, or MWE lexicons. MWE LRs especially comprise several types that will be reviewed in-depth in section 2.4.

One of the earliest lexical database was the Longman Dictionary of Contemporary English (1978) which utilised computational methods to build an easy access language dictionary. The target end-users were humans, so the importance of readability features was considered in the design. However, several human-oriented lexical resources have recently been used for NLP purposes and vice versa. Granger and Paquot (2012, p. 3) state this is because 'the line between these two types of lexical resources is progressively narrowing, and NLP resources like WordNet are increasingly being integrated into human-oriented tools.'. However, the primary distinction between human and machine oriented LR is related to the representation of linguistic information, where more strict formal representations are preferred to eliminate the language ambiguity caused by data noise from models with loose formalisms.

Regarding types of developer, lexical knowledge bases can be categorised in two ways. First, they can be expert-built (e.g., Wordnets, Framenets, Verbnets) where a designated expert or a group of specialists build high-quality lexical resources. Second, they may comprise collaborative LR in which many contributors, mostly non-experts, build a, LR using the advantages of crowdsourcing tools (e.g., Wikipedia, Wiktionary, Omegawiki).

3.3.4.3 Linguistic annotation

In the development of an AMWE lexicon, several layers of representations are added to each lexical entry to enhance its usability in various applications. These metadata

include what can be classified as a type of linguistic annotation. This term will be discussed briefly in this section. However, an early definition of linguistic annotation was provided by Leech (1997 p. 2) who defined it as:

> *'the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process: the linguistic symbols which are attached to, linked with or interspersed with the electronic representation of the language materials itself'.*

Since the 1980s annotation has become an increasingly active research area that involves the enhancement of LR with linguistic data to improve computational representations and discover new insights from language data through various levels of linguistic representation. The Lancaster-Oslo-Bergen (LOB) corpus of English was the first available corpus with automatic morpho-syntactic annotation (Garside et al., 1988). This was followed by the building of the first English treebank based on a one million news corpus (Marcus et al., 1993) which was later extended to include multiple languages, including the Penn Arabic Treebank (Maamouri et al., 2004). The British national corpus was developed and enhanced with different types of linguistic annotation and is considered the first available large-scale annotated corpus to be used intensively in most corpus-based research (Clear, 1993). Leech (1997) describes three main areas where annotation can play an important role: extracting information, LR reusability, and multi-functionality of annotated LRs. Based on Leech's (1997) inventory of annotation layers, Rayson (2002 pp. 19–21) lists with examples the possible linguistic annotation layers which include the following 13 levels of representation: orthographic, phonological, phonetic, morphological, lemma, prosodic, grammatical, syntactic, semantic, discoursal, pragmatic, stylistic, and application-oriented annotations. Therefore, in the lexicon for the current research, different types of annotation will be considered and these will incorporate the most relevant layers of annotation in the representational model for an AMWE lexicon. Thus, the use of this term in the thesis refers to multiple layers of linguistic annotation

designed to enhance the quality, functionality, and reusability of the developed AMWE LR.[36]

Regarding annotation tools, several surveys have been conducted in the literature to evaluate the available tools based on a list of criteria that encompasses tool features and methods for tackling different annotation tasks. For instance, Biemann et al. (2017) critically evaluate state-of-the-art existing Collaborative Web-Based Tools for Multi-layer Text Annotation based on 20 criteria. They find it hard to determine the best existing annotation tool as each tool or system has its strong and weak points, which makes them more suitable for specific annotation tasks.

3.3.4.4 MWE Computational processing

MWE computational processing in this thesis refers to the main computational tasks that pose different kinds of challenges in MWE research and include the following:[37]

MWE extraction: this means finding various computational techniques for discovering MWE items from different types of language data to create a new LR or enhance existing LRs.

MWE LR representations: this includes building different types of MWE lexicons and the enhancement of these LRs by linguistic annotation and a computational representational system that in turn improves LR re-useability and provides them with multifunctionality.

MWE identification: this means annotating MWE in the running text, the result of which is annotated text with MWE labels.

Embedding MWE knowledge into practical applications: this task includes the effort needed to designNLP tools that take advantage of MWE LR, such as tokenisation, language parsing, MT information retrieval, and sentiment analysis .

---

[36] Several books have been devoted to covering various aspects of linguistic annotation and these can be referred to for more in-depth detail on the science of annotation (e.g., Leech, 1997; Fort, 2016; Lu, 2014; Ide and Pustejovsky, 2017; Ho-Dac, 2009)

[37] This is not intended to be an exhaustive list of all computational tasks that include MWE treatment. Other tasks can also be found in the literature such as MWE interpretation and disambiguation.

However, there is no sharp division between these three MWE computational tasks as in many cases they overlap intensively, especially MWE extraction and identification. Such strong interactions between MWE tasks justifies the equivocal boundaries that are described as existing between them in the literature. Figure 3.3 illustrates the relations between several MWE processing tasks and is adopted from Constant et al. (2017, p. 843).

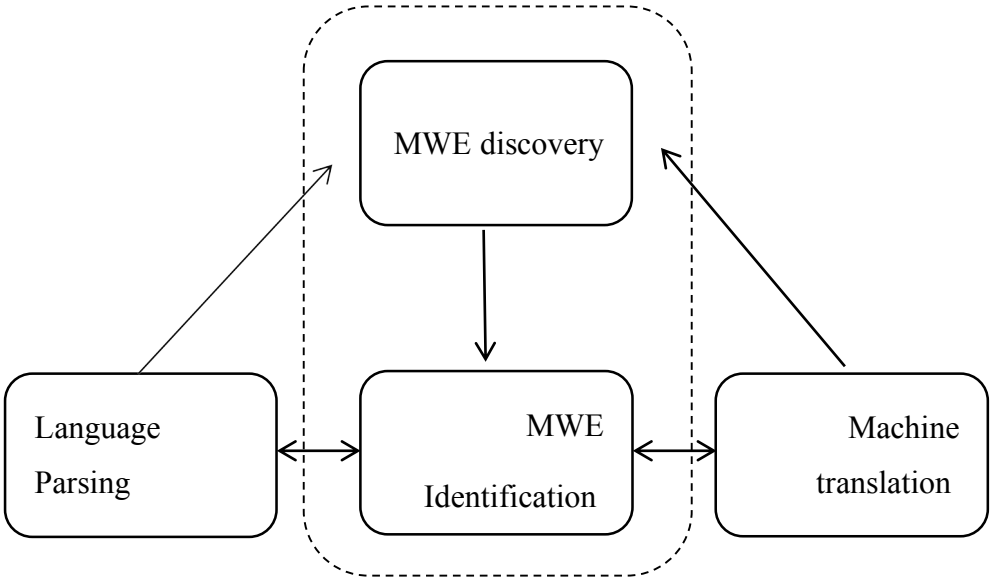


Figure 3.3: Interactions between MWE processing tasks with two examples of two NLP tasks.

As shown, each MWE task or application has a positive effect on the other as illustrated by the direction of the arrows. Thus, discovery improves identification and MT, and language parsing can be used in various MWE extraction models. Moreover, identifications and NLP tasks have a bidirectional effect which means there are supportive relations between them. Building and enhancing MWE LRs can be added to this figure to represent another significant area of research in MWE processing tasks. Nevertheless, for practical reasons, his conceptual framework was adopted in the thesis to delineate the boundaries between multiple MWE computational tasks. This concern in this project is on the first two tasks relating to MWE extraction or discovery and building MWE LRs where the aim is to experiment with several MWE extraction models to build a new MWE lexicon with an intensive computational formalism that can be used in future work to enhance other MWE computational tasks.

Chapter 2 presents an in-depth survey of MWE extraction methods and MWE representations.

3.3.4.5 NLP tasks or applications

The core objective of NLP tasks and applications is to facilitate an understanding of natural language and reduce different types of ambiguity in languages. However, because MWE is the source of a significant amount of ambiguity in language, adequate computational processing will improve nearly all NLP tasks and applications. Figure 3.4 presents a list of NLP applications in which MWE knowledge can be integrated.

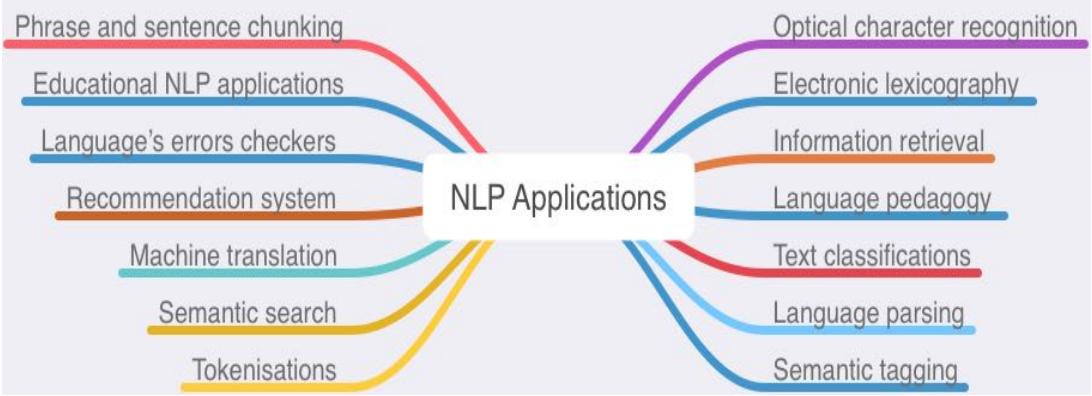


Figure 3.4: Examples of NLP applications in which MWE can be integrated.

For instance, in MT, which is where most application-oriented MWE research in the literature can be found, several studies conclude that integrating MWE LRs into the translation process considerably improves the system output (e.g., Tan and Pal, 2014; Monti, 2015; Lambert and Banchs, 2005; Ren et al., 2009; Pal et al., 2010; Carpuat and Diab, 2010a). Rikters and Bojar (2017b) examined the impact of MWE information on the statistical bilingual n-grams model of MT between English and Spanish and vice versa. They found there to be a substantial improvement and that the more MWE data was integrated into the MT model the better the quality of the translation output.

3.4 AMWE properties

In the following subsections, the core linguistic features of AMWE will be briefly illustrated. This is an essential step in understanding the behaviour of AMWEs in their various linguistic manifestations. Rayson et al. (2010, p. 2) stress that 'in order to

develop more efficient algorithms, we need a deeper understanding of the structural and semantic properties of MWEs, such as morpho-syntactic patterns, semantic compositionality, semantic behaviour in different contexts, cross-lingual transformation of MWE properties.' However, it is important to note that AMWE examples are used in most cases because the objective is to demonstrate their various linguistic properties.

3.4.1 Arbitrarily prominent co-occurrence

Nearly all definitions of MWE in the literature concentrate on this core MWE property which is considered a type of statistical idiomaticity, as illustrated in Baldwin and Kim (2010), within the adopted concept of MWE in this thesis. This is because frequency-based data on the co-occurrence of words is one of the most reliable and consistent objective criteria for identifying MWEs in running text. These types of lexical unity are often illustrated by the term 'collocations'. One of the earliest definitions of this phenomenon by Firth (1961 p. 181) highlights this characterisation:

> *'Collocations of a given word are statements of the habitual and customary places of that word.'*

Subsequent definitions also consider this criterion to be the best predictor for this type of MWE. For instance, Bartsch (2004, p. 76) defines collocation as:

> *'lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other.'*

Hence most MWEs in languages tend to have significant high frequency and consist of adjacent words. This can be seen in various examples of AMWEs, as illustrated in the nominal (1) and prepositional (2) genitive phrases below, where they are always fixed in their structures to these specific words in SA. Although there are several alternatives words with a similar meaning, they do not reach a frequency of co-occurrence high enough to confer the status of institutionalisation or permanency.

خالي الوفاض

1- ḫālī alwafāḍ.

Lit. Empty of pond.

Idi. Useless or ignonant.

على قدم وساق

2- ʾalā qadam wasāq.

Lit. On foot and leg.

Idi. By leaps and bounds.

A multitude of examples can also be found in English, where corpus linguistic studies have identified a list of various types of formulaic frames that seem to occur frequently. For example, Hunston and Francis (2000) found that the word 'matter' in English was usually found in the frame 'a matter of V-ing'. Wray, (2002a) explained that, in language processing mechanisms, there is a significant correlation between high-frequency phrases and the likelihood of being MWEs. She stated that 'the more often a string is needed, the more likely it is to be stored in prefabricated form to save processing effort, and once it is so stored, the more likely it is to be the preferred choice when that message needs to be expressed' (p. 25). Several studies have found a strong link between the high frequency of sequences and the holistic processing of human languages. For instance, using an eye-tracking paradigm, Underwood et al. (2004) identified an advantage for native speakers regarding the processing of MWEs. Durrant (2008) also found a significant relationship between high frequency of occurrence and the mental representation of lexical items in a series of lexical decision experiments conducted with adult second language learners.

Other advantages can be observed in the use of frequency and computational search methods in detecting MWEs, such as consistency and the capacity to handle a significant amount of data in a matter of seconds. Furthermore, statistical evidence in the form of frequencies and probabilistic data, unlike intuition, offers a rich description of the authentic usage of language phenomena and helps distinguish between high and low expressions in the targeted language. Thus, frequency data is essential in the process of extracting MWEs. Sinclair and Renouf (1988) also emphasised that "no description of usage should be innocent of frequency information" (p. 152).

However, relying entirely on frequency and statistical data might mislead the researcher due to their inherent limitations. For instance, frequency data cannot differentiate between figurative or literal phrases. Another problem encountered by researchers in Arabic is that some words and phrases have entirely different meanings depending on pragmatic and semantic contexts, which often reduces the reliability of frequency information. Thus, the researcher should be aware of the pitfalls of complete dependency on frequency data and should make use of other supplementary criteria when discovering AMWE items.

3.4.2 Discontinuity in AMWEs

Flexible MWEs permit a variety of words or phrases to be inserted between components of their core lexemes at various degrees of intervention. This property of discontinuous construction poses various problems in computational processing. Researchers must find different ways to overcome these challenges, particularly in MWE discovery and identification tasks. For instance, the verbal AMWE example in Table 3.9 shows multiple types of intervention between the verb *ʾatā* and its object, the prepositional phrase *ʿalā alʾaḫḍar walyābis*, as can be seen in example (1) which shows one-word insertion, example (2) which shows adjectival phrase insertion, and example (3) which shows another verbal phrase insertion, all of which are discontinuous AMWE.

Table 3.9: Examples of contiguous and discontinuous AMWEs.

Contiguous AMWE	Discontinuous AMWE
أتى على الأخضر واليابس	أتت الحرب على الأخضر واليابس
ʾatā ʿalā alʾaḫḍar walyābis.	ʾatat ʿalḥarb ʿalā alʾaḫḍar walyābis.
Lit. It came on the green and dry.	Lit. the war came on the green and dry.
Idi. It destroyed everything.	Idi. the war destroyed everything.
	أتت الحرب الظالمة على الأخضر واليابس
	ʾatat alḥarb aḍḍālima ʿalā alʾaḫḍar walyābis
	Lit. the injust war came on the green and dry.
	Idi. the injustice war destroed everything.
	أتى تفشي الفساد على الأخضر واليابس
	ʾatā tafaššī alfasād ʿalā alʾaḫḍar walyābis.
	Lit. came spread of corruption on the green and dry.
	Idi. widespread corruption destroy everything.

These various arbitrary modifications of MWEs should be considered carefully in MWE processing tasks. For instance, at the tokenisation and POS tagging levels, the discovery methods should find a way of capturing flexible AMWE. Thus, these challenges can be addressed with an appropriate morphosyntactic analysis that eliminates the ambiguities in language parsing outputs by accommodating discontinuous MWEs and distinguishing them from fixed MWEs.

3.4.3 Non-compositionality

Non-compositionality is the core, common semantic feature of most types of AMWE; it is primarily observed when the meaning of MWE cannot be directly derived from the meaning of its component parts. This semantic characterisation of MWEs ensures these types of phrases stand out in NLP research because they produce diverse types of semantic ambiguity in the generation and understanding of natural language. However, not all MWEs have the same degree of non-compositionality as strong variability can sometimes be observed. MWEs with a high degree of non-compositionality are mostly described by the term ‘idioms’ in the literature and can be distinguished by non-literal translation and non-substitutability. These are two popular methods used to discover these types of opaque phrases. The former means that MWEs of this type cannot be translated with the exact meaning as a sequence of words, but instead have to be mapped to an equivalent single word or phrases in the

corresponding language to achieve an adequate translation output. The latter means this type of MWE usually tends to be a fixed construction, particularly at the lexical level; hence the core lexemes of non-compositional MWEs cannot be substituted with other, similar lexical items. These two manifestations of semantic non-compositionality can be seen in the following examples of AMWE:

Table 3.10: Examples of lexically fossilised AMWEs.

بادئ ذي بدء	ابن حلال
bādiʾ ḏī bidʾ	ibn ḥalāl
Lit. start of the start	Lit. son of halal
Idi. First of all	Idi. A respected gentleman
بين يوم وليلة	حبر على ورق
bayn yawm walayla	ḥibr ʿalā waraq
Lit.between day and night	Lit. ink on paper
Idi. In a quick manner	Idi. Impracticable (plan, etc.)
حاطب ليل	عن بكرة أبيهم
ḥāṭib layl	ʿan bakrat ʾabīhim
Lit. a night woodchopper	Lit. *Riding* their father's *camel*
Idi. An unreliable person.	Idi. altogether

These have a high degree of non-compositionality; thus, we cannot find a correct literal translation for them in English and their constituent components cannot be replaced with other substitute lexical items in SA. In the following subsections discussing MWE properties, more examples of this type of lexical unit will be presented.

3.4.4 Ambiguity

Due to the distinctive linguistic features of MWE, ambiguity[38] can be seen in AMWEs at various levels of linguistic analysis. At the orthographic level, several MWEs may be classified incorrectly as one-lexeme words because SA is characterised by highly

[38] In linguistics, a distinction is drawn between ambiguity and the complementary term 'vagueness' where the former means the type of ambiguity that can be resolved or represented by human or syntactic analysis while the latter refers to the type of ambiguity that cannot be resolved or represented in a systematic way (Bussmann, 2006).

ambiguous agglutination in which orthographic strings might consist of up to five syntactic units (section 3.2.1.3). The following examples show the type of orthographical ambiguity that can be observed in AMWEs. Such AMWEs are more than syntactic tokens despite being written as a one space-delimited string.

Table 3.11: Examples of one string type AMWEs.

بحرارة	بحذافيره
bi.ḥarāra	bi.ḥaḏāfīri.hi
Lit. with hotly	Lit. from all sides
Idi. warmly	Idi.in an exact manner of something.
بصدد	برمته
bi.ṣadad	bi.rummati.hi
Lit. in front of	Lit. with his neckband
Idi. regarding	Idi. entirely.

Another type of ambiguity, derived from the semantic analysis of MWEs, occurs when the system reads and has to decide whether a sequence of words should be yielded as MWE. This discrimination of multiple reading interpretations is necessary because, based on the context, several MWEs might be used in terms of either their literal or idiomatic meaning, as can be seen in this AMWE.

رأيت عين الرجل

raʿytu ʾayn arrajul

Lit. I saw the man's eye.

Idi. I saw the man himself.

This type of ambiguity can also be seen in many MWEs in English. For instance, the phrase, *by the way*, depending on its context can be used either in terms of its literal or as in most cases its figurative meaning. In the linguistic literature, an enormous amount of research has been devoted to this type of semantic ambiguity, which is known as polysemy[39]. Semantic ambiguity poses diverse challenges to adequate

[39] Polysemy and homonymy two are terms for describing semantic ambiguity in linguistics. However, polysemy is used ‘when an expression has two or more definitions with some common features that are usually derived from a single basic meaning’ and ‘The distinction between polysemy and homonymy cannot be drawn precisely’(Bussmann, 2006, p. 918).

reading at the word and phrase levels, particularly for computational methods because they do not have parallel techniques for accessing the context, intonation, or situational information that is available for use in human communication. However, to avoid repetition, various types of ambiguity in AMWEs are illustrated where appropriate when considering different types of linguistic variability of AMWEs in section 3.4.5.

3.4.5 Variability in AMWEs

One of the most common features of MWEs is instability and variation these linguistic units allow at different linguistic levels, which requires a comprehensive analysis of MWEs based on representative samples of authentic usage. In his analysis of English idioms, Langlotz (2006) listed diverse types of variations evident at different linguistic levels, such as institutionalised, usual, and occasional variants. The first of these relates to stable alternation and accrues to phrases which lead to the institutionalised status of idioms in the language. Regarding the second and third variants, a *Usual* variant is a variant form that frequently occurs in the phrases while a *Occasional* variant is the opposite. However, most of these types of variants can be found in AMWEs; for instance, the alternate support verb phrase (ʾaḫaḏa zimām almubādara أخذ زمام المبادرة) which initially changes from the original phrases (ʾaḫaḏa almubādara أخذ المبادرة) has an institutionalised *variant* because the corpus evidence indicates high-frequency use of the first phrase in the actual use of SA. The following subsections briefly illustrate with examples the variability in AMWEs at lexical, morphological, syntactic, and semantic levels.

3.4.5.1 Lexical

An analysis of several MWEs shows that most AMWEs allow some substitutions in their lexical items, although the underlying meaning of the MWE is preserved. It is also evident that lexical variations can be explained by the semantic characteristics of MWEs in different contexts. Fellbaum (2007) stated that, "lexical selection is even stronger in expressions that are not semantically transparent" (p. 9). The lexical substitutions of MWEs vary in terms of the frequency of occurrences, based on different situations and kinds of discourse. Many instances of different types of lexical

variation can be noted. Examples of variations in verb, noun, adjective and prepositions variations are presented in table 3.12 with examples of AMWEs.

Table 3.12: Distinctive types of lexical variations in AMWEs.

lexical variation	AMWEs example
Verb	تجمد/وقف/الدم في عروقه
	tajammad/waqaf/addamu fī ʿurūqihi
Noun	قيد أنملة / شعرة
	qayda ʾunmula / šaʿra
Adjective	قضية محسومة / منتهية
	qaḍiyya maḥsūma / muntahiya
Prepositions	على/ ب/ الرغم من
	ʿalā / bi arraġmi min

The lexical variation in these examples has no substantial impact on the meaning, which means these phrases have the same meaning despite their multiple lexical variants. These types of lexical flexibility are considered in the representational model of the AMWE lexicon used in this research because this will enhance the multifunctional use of the developed LR in various potential applications. The inclusion of lexical variations also assists in AMWE identification tasks which allow the recognition of several AMWE variants.

3.4.5.2 Morphological Variation

One of the most notable features characterising Semitic languages is the inter-digitation of many morphological forms of words that are derived from one root. This explains the core meaning of all its derivational and inflectional forms; thus, words in Arabic cannot be analysed directly by the concatenation of morphemes as they require a more comprehensive analysis of various word patterns (section 3.2.1.2). The rich morphological nature of words results in various types of derivational and inflectional forms of MWEs that should be reflected in MWE processing tasks. Table 3.13 presents different examples of morphological variant types in AMWEs.

Table 3.13: Examples of morphological variation in AMWEs.

Morphological variation	MWEs example
Tense and person	كظم /يكظم /تكظم غيظه
	kaḏam /yakḏum /takḏum ġayḍa
Number	السُّوق / الْأَسْوَاق السَّوْدَاء
	assūq / alʾaswāq assawdāʾ
Gender	شَاعِر / شَاعِرَة مَطْبُوع / مَطْبُوعَة
	šāʿir / šāʿira maṭbūʿ / maṭbūʿa

These examples illustrate the main types of morphological inflections in MWEs which includes tense, person, number, and gender, and the words usually inflect based on the agreement rules of the SA syntactic system mentioned briefly in section 3.2.1.4. The first phrases show three tense and person inflections of the verb *kaḏam /yakḏum /takḏum*. The context usually determines the right inflected forms in these morphological variants. This rich morphology requires extensive attention to reduce the noise data in MWE processing using different computational methods such as stemming, lemmatisation, and morphological disambiguation. Additionally, a proper representation schema also should take account of all the morphological variation potentials to extend its coverage to all inflectional and derivational forms of AMWEs.

3.4.5.3 Grammatical and Syntactic Behaviour

The grammatical and syntactic behaviour of AMWEs reveal various types of variability in the syntactic structures and grammatical variables. Most MWE structures in the literature can be found in SA. Table 3.14 presents various syntactic structures of AMWEs with an analysis of their grammatical function.

Table 3.14: Examples of common AMWE syntactic patterns.

Syntactic Structure	Grammatical function	Examples
noun-adjective	[nominative subject-adjective 'attribute']	السَّوادُ الأعْظَم assawādu alʾaʿḏam Vast majority
verb-noun-pronoun	[nominative subject-object-complement]	عيلَ صَبْري ʿīla ṣabrī Fed up
noun-noun	[nominative subject- genitive noun]	غَريبُ الأَطْوار ġarību alʾaṭwār Changeable of mind
noun-adverb-noun	[particular-genitive adverb-genitive noun	عَلَى حِينِ غِرَّة ʿalā ḥīni ġirra Suddenly
preposition-noun	[particular- genitive noun]	عَلَى الفَوْر ʿalā alfawr Immediately

All AMWE structures can be mapped onto the traditional classifications of SA sentences. These include nominal, verbal, and other types of sentence[40]that include structures beginning with other word classes (e.g., preposition, adverb, adjective). In the following quotation, Holes (2004) provides an overview of the grammatical structure of the SA sentence that helps in understanding various syntactic manifestations in AMWE:

> *'Syntactically speaking, a sentence in written Arabic consists of a subject and predicate. The subject may be free standing, that is, a noun/independent pronoun; or dependent, that is, consisting of one or more bound morphemes that form part of the verb if there is one and that indicate the person, number, and gender of the subject. The predicate may or may not contain a verb. If it does contain one, the subject may or may*

[40] This type of structure is called a 'semi-sentence' by traditional grammarians in Arabic. It also has specific implications for grammatical functions in SA.

not be free standing; if it does not, the sentence subject must be free standing. The verb may or may not have a complement' (p. 251).

Regarding the grammatical functions of constituents, one of the most notable properties of the grammatical behaviour of AMWEs is that they usually allow for changes in the constituent order. For instance, the word order of the second example in the previous table can change from [ʿīla ṣabrī عيل صبري] to[ṣabrī ʿīla صبري عيل] without any impact on its core meaning (section 3.2.1.4).

However, having considered all these types of variation in the syntactic and grammatical behaviour of MWEs, it is vital to take account of all these phenomena in the current research, specifically in the development of comprehensive standards and formalism for AMWEs.

3.4.5.4 Semantic and pragmatic analysis

The semantic and pragmatic analysis of the behaviour of AMWEs reveals several phenomena that can be observed in various types of AMWE. In corpus-based research on Arabic idioms, Abdou (2011, p. 222) found five main patterns of semantic extensions based on the meaning and authentic usage of AMWEs, which are as follows; 'metaphor, metonymy, interaction of metaphor and metonymy, and semantic extension based on conventional knowledge, hyperbole, and emblematising'. In addition, based on a comprehensive corpus-based analysis, he also found that prepositional phrases in Arabic were more commonly used figuratively than other syntactic structures. Furthermore, the semantic analysis of AMWEs shows that they can represent a range of well-known semantic fields, such as social relations, wishing and cursing, and discourse markers. In this research a semantic lexicon will be built.

Thus, different classifications of semantic fields will be considered and semantic labels added for each of the AMWE lexical entries. A semantic lexicon developed for English has been found to be very useful in various NLP semantic based applications such as semantic tagging and concept-based search tools. For instance, the semantic analysis system (USAS) developed by Rayson et al. (2004) paved the way for many subsequent projects in English and other languages which included building a semantic tagger for other languages and the enhancement of these taggers by creating different types of semantic lexicon (e.g., El-haj and Rayson, 2016; Löfberg et al.,

2005; Piao et al., 2006; Rayson, 2008; Piao et al.,, 2015; Piao et al., 2017; El-haj et al., 2017).

The discursive behaviour of MWEs shows they are used for different pragmatic purposes. Therefore, knowing these different discursive functions, semantic fields, and the relations between MWEs in different contexts plays a significant role in the semantic and pragmatic applications of an AMWE lexicon. Several classifications have been proposed in the literature; for instance, Moon (1998) classified the text functions of MWEs into five main semantic fields, as shown in Figure 3.5 along with their English examples, which are phrases that clarify the meaning of these categories.

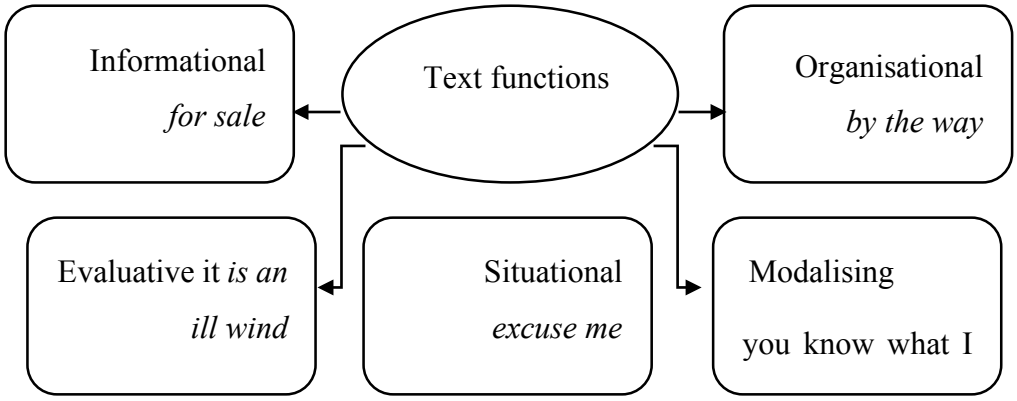


Figure 3.5: Text function categories (Moon, 1998).

However, in Arabic, similar examples can be found in all these text function categories; for instance, the phrase (على سبيل المثال *ʿalā sabīl almiṯāl,* for example) is used to organise the text, and the phrase (للبيع *lilbayʿ* for sale) is used to express an informational function in the discourse.

The current research is based on the analysis of MWE data, conducted to develop an intensive typology model for the semantic and pragmatic functions of AMWEs which describes their behaviour in detail and shows the most frequent and essential phrases that can be used to express various meanings in different discourses. Taking account of all these linguistic features of AMWEs is very important in developing semantic LRs that can be utilised in multiple content-based applications.

3.5 Typology of multiword expressions

This section provides a brief review of the most influential classifications of MWEs that have been proposed in the literature. Due to the heterogeneous nature of the MWEs and the fuzzy borders of research areas concerning this phenomenon, several typologies have been suggested and implemented from different linguistic perspectives. For instance, lexicographically oriented classifications (e.g. Moon, 1998; Cowie, 2001) and a typology for pedagogical purposes (Nattinger and DeCarrico, 1992; Lewis and Conzett, 2000; Lewis and Gough, 1997), psycholinguistic classifications (Wray and Perkins, 2000; Sidtis, 2011; Wray, 2002a) and other classifications suggested from NLP perspectives (Tschichold, 2000; Meghawry et al., 2015; Diab and Krishna, 2009; Sag et al., 2002; Ramisch, 2015a) However, most of these classifications were based upon the principle linguistic features of MWEs that include syntactic structures, flexibility and fixedness of the phrases, semantic level of non-compositionality, or the discourse function.

3.5.1 Fillmore et al.'s typology

An early typology involving MWEs suggested by Fillmore et al. (1988) from grammatical construction perspectives classified idiomatic expressions into three main categories based on the familiarity of the lexical items in expressions and the mode of combination between them.

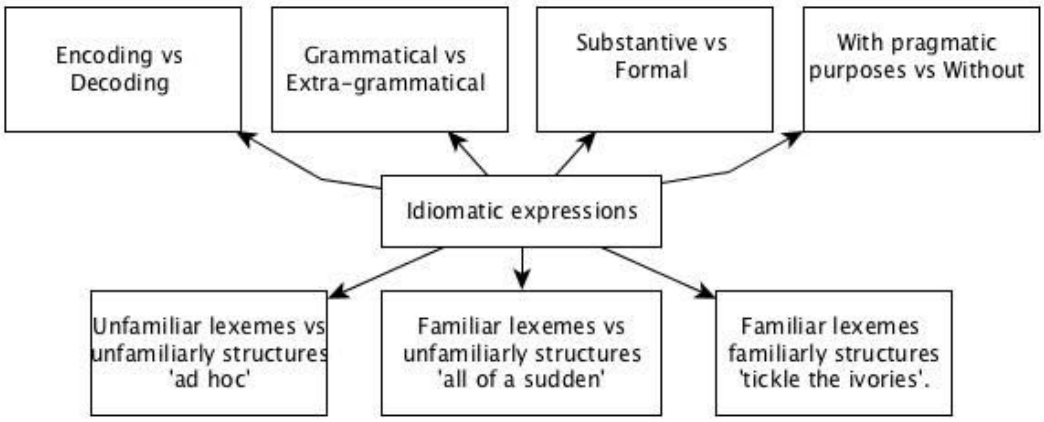


Figure 3.6: A typology of idiomatic expressions (Fillmore et al., 1988, p. 506).

Other perspectives classify these expressions into eight classes based on their various linguistic features, as can be seen in Figure 3.6. The first category, decoding

expressions, refers to the type of MWE whose meaning cannot be understood without previous experience of the meaning and the correct use in context while the second category includes expressions that can be interpreted using prior knowledge.

The grammatical phrases include the conventional syntactic constructions that might be used with an idiomatic meaning such as phrasal verbs in English. In contrast, the term 'extra-grammatical' is used to refer to MWEs that have unique syntactic structures that contravene most grammatical rules in English phrases such as *all of a sudden* or *by and large*. Formal MWEs means expressions that can be used as a template, such as *lexically open idioms* which include the corresponding category, and substantive idioms which include all the *lexically filled idioms* under the formal expressions class. This can be seen in the popular formal idiom '*the_x_er the_y_er'* which includes many substantive examples such as *'the bigger, the better'*.

The final two categories distinguish idioms based on their pragmatic use, as some expressions in languages are associated with a specific pragmatic uses while others are free from these constraints. The former includes expressions such as *good morning* and *what's up?* while the latter includes phrases such as *all of a sudden* and *by and large*. However, all these categories can be found extensively in SA; therefore, in the typology proposed in this thesis, all these possibilities for classifying AMWEs will be considered based on their contextual and linguistic features, as will be discussed in depth in chapter 7.

3.5.2 Cowie's typology

In English, the classification by Cowie 1998, 2001) concentrated on the semantic properties of MWEs. Figure 3.7 summarises Cowie's typology which divided word combinations into two main categories, composites and formulae. The composites were then subdivided into restricted collocations, figurative idioms, and pure idioms, while the formula was subdivided into two classifications; routine formulae and speech formulae.

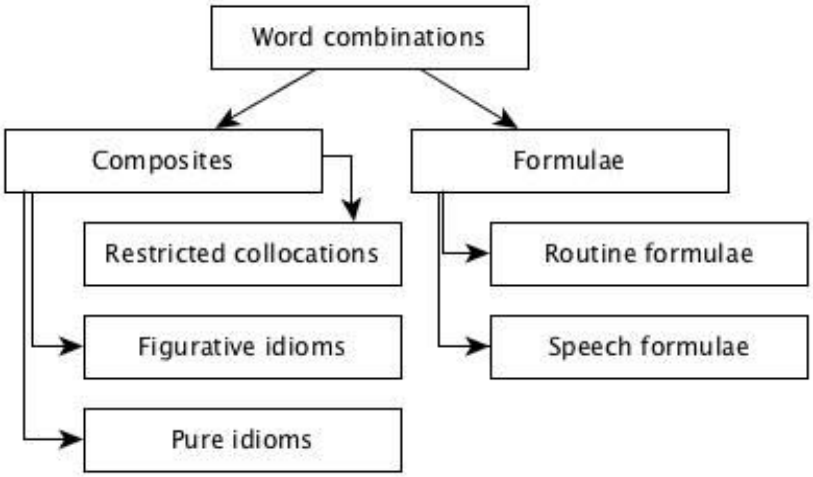


Figure 3.7: Cowie's classification of word combinations.

3.5.3 Mel'čuk's typology

Another important MWE classification was suggested by Mel'ćuk (1998, 2003, 2012) in his work on Meaning-Text theory. The typology is very similar to the work of Cowie with some changes in the terminology. Figure 3.8 presents Mel'čuk's classification of word combinations.

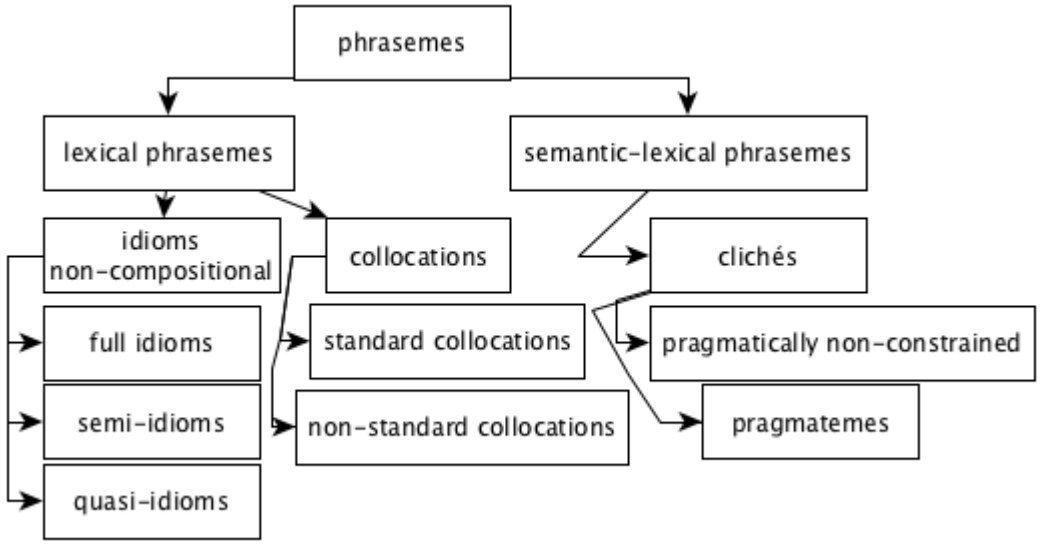


Figure 3.8: Classification of phrasemes according to (Mel'Čuk, 2012 p. 42).

This comprehensive typology uses three main classes in representing MWEs or phrasemes, which are idioms, collocations and clichés. The first class includes non-compositional expressions with various degree of semantic opacity and the second

one means semantically compositional phrases like the support verbs in English. The last class, which is also called 'semantic-lexical phrasemes' or 'lexical anchors', covers multiple types of compositional expressions that are used for specific communicative situations such as 'Happy birthday to you', 'no matter what' and 'no parking'. In SA equivalent expressions can be found which represent all the classes mentioned above of MWEs based on Mel'Čuk's typology.

3.5.4 Burger's typology

Another classification proposed by Burger (2007), Burger et al. (2002), Burger and Sloane (2004) and Wray (2012) concentrated on the practical use of the phrase in different discursive contexts. Thus, the phrases were classified according to their various functions in discourse, as can be seen in Figure 3.9. Burger categorised what he termed the phraseological units into three main groups; referential, structural and communicative. At the second level of classification, the referential units were subdivided into nominative and propositional phraseological units.

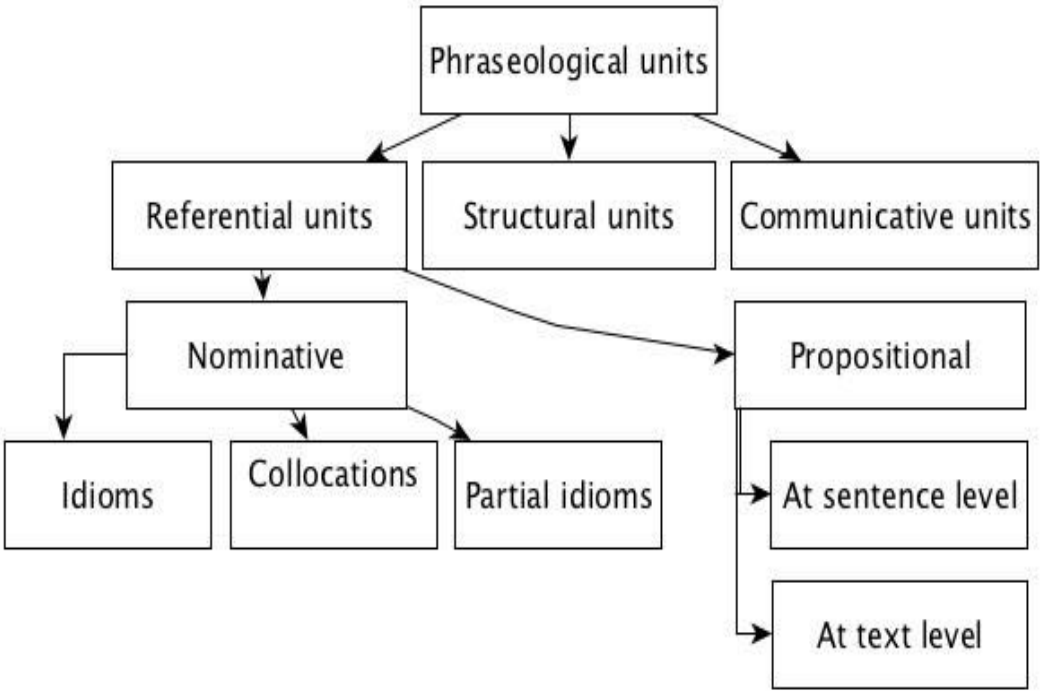


Figure 3.9: Burger's typology of phraseological units.

3.5.5 Sag et al. 's typology

From NLP perspectives, several studies have presented various typologies of MWEs that reflect the different procedures and experimental settings used, particularly the distributional frequency-based approach to collocation extraction (e.g. n-gram model and AMs). An example of the classification of word combination can be seen in the work of Sag et al. (2002) who outlined two main categories of lexicalised and institutionalised phrases among MWEs. Figure 3.10 summarises the main classifications of word combinations.

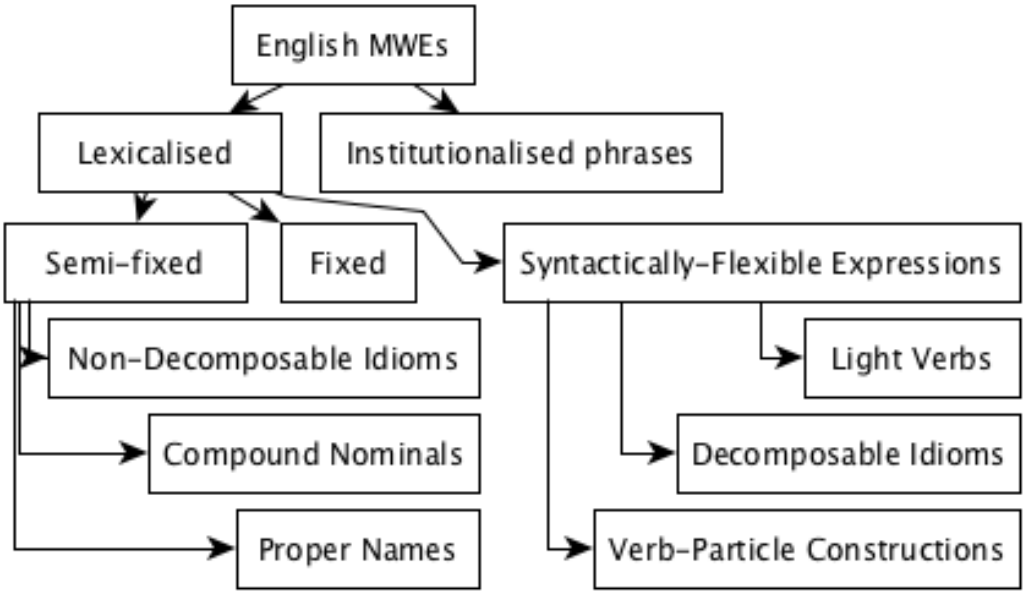


Figure 3.10: Typology of English MWEs (Sag et al., 2002).

3.5.6 Ramisch's typology

Another typology of MWEs was suggested by Ramisch (2015a) in the context of building a framework of MWE acquisition. The classifications were based on the previously mentioned MWE typologies with a specific focus on the morphosyntactic role of MWEs in the sentence and the difficulty of expressions in the computational treatment.

As shown in Figure 3.11, MWEs in this typology are divided into six main classes, each of which might include other detailed subclasses, such as verbal MWEs which includes phrasal verbs and light verbs.

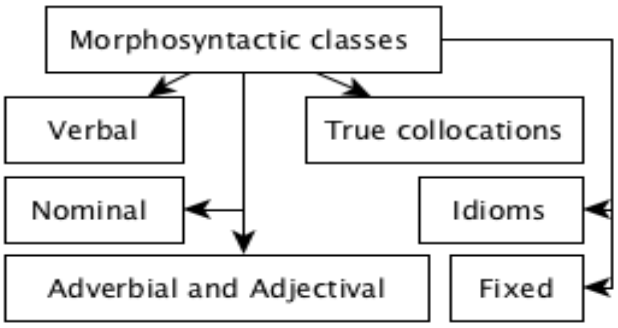


Figure 3.11: MWE types (Ramisch, 2015a, pp. 42–44).

In Arabic, several classifications of word combinations have been suggested based on their linguistic characterisations, as described in section 2.4.5. In this research, the typologies of MWEs mentioned previously will be analysed and their feasibility in SA assessed to develop an AMWE classification which represents the main types of AMWE described in section 2.5.6.

3.5.7 Adopted Typology of AMWE

Rather than following an elaborate typology of AMWE which might pose various problems in extraction and evaluation tasks, a simplified classification adopted from Ramisch (2015) will be followed with several modifications, especially in the sub-classifications, to suit the linguistic properties of SA. The main advantage of this classification is that it is flexible and scalable; thus, in the adopted concept of AMWE (section 3.3.2) the research includes a range of AMWE types that are not restricted to specific syntactic or semantic category.

This typology is based solely on the morphosyntactic heads of AMWE sequences which could theoretically cover most AMWE structures in SA. However, for practical reasons, several constraints will be imposed at the extraction stage in sub-classifications due to the scale limitations of the current research. Figure 3.12 shows the main categories of the AMWE typology. A more detailed description and examples are provided in section 4.6.

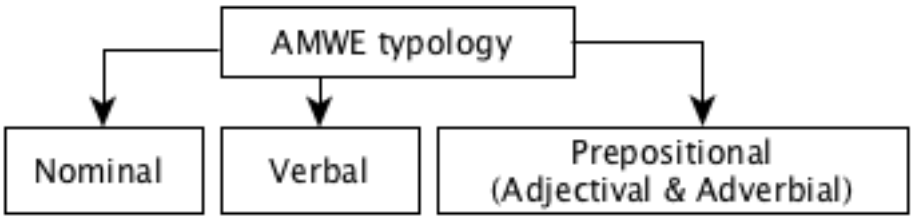


Figure 3.12: The typology of AMWEs based on the head class of the phrase adopted from Ramisch (2015).

In the current research, three major categories of AMWE, prepositional, adjectival and adverbial, were combined into one class of AMWE because of the limited number of these expressions in the language data.

3.6 Summary

The heterogeneous nature of AMWEs can be observed at all linguistic levels, particularly in morphologically rich languages such as SA. These various linguistic features render most MWE processing tasks challenging. An in-depth understanding of several related linguistic phenomena is required to improve the computational treatment of AMWEs and eliminate the language ambiguity caused by inadequate treatment of this complex phenomenon. In this chapter, a brief theoretical background on SA and its core linguistic properties has been presented, followed by a brief description of the core concepts used in this thesis and several issues related to the terminology.

The distinctive characterisation of AMWEs at various linguistic analysis was then described, followed by a review of existing typologies of MWEs in the literature. This review of related work on MWEs clarifies the nature of the linguistic processing and analysis that will be presented in chapters 4 to 7.

4 A Hybrid model for Constructing AMWE reference data

4.1 Introduction

Reference or gold standard data play a significant role in building a high-quality MWE lexicon and the evaluation of various MWE-aware NLP tasks, especially MWE automatic extraction models as described in research presented in section 3.2.4.3. However, as revealed in the survey on existing AMWE LRs (section 3.3), no large scale and well-validated machine readable AMWE lexicon exists that can be adopted and used as reference data in AMWE computational tasks.

Another point to mention relates to the fact that the reference AMWE lists constructed in the series of experiments reported in this chapter are aligned with the established conceptual framework of this complex phenomenon in SA, which was described in detail in chapter 3 and specifically in sections 3.3 and 3.4. Thus, in this chapter, the development of several AMWE data sets will be reported which can be used as reference data in the empirical evaluation of further AMWE extraction experiments reported in chapters 5 and 6. Furthermore, the AMWE lists used in the process of building a large scale computational AMWE lexicon can ultimately be used as high-quality AMWE LR in various NLP applications.

In the extraction methodology for the experiments, a hybrid approach was adopted that exploited the statistical and linguistic methods described in section 2.2. Hence, the development of reference data was based on various linguistic and statistical techniques implemented in the semi-automatic model for extracting multiple types of AMWEs from a large SA corpus.

This chapter is organised as follows. Section 4.2 presents brief extraction guidelines and the recommendations observed in the experiments. Sections 4.3 provide details about the corpus used in this study. In section 4.4, a brief explanation is provided of the automatic linguistic toolkits used in the AMWE extraction model. Section 4.5

describe the methodology adopted in the experiments and the central processing components involved in the extraction architecture.

Section 4.6 presents the original work by reporting a series of experiments for extracting various constructions of AMWE candidates which are then validated and evaluated in section 4.7. The final two sections, 4.8 and 4.9, discuss the findings of the experiments and summarise the overall results. They also present an introduction to the extension AMWE experiments conducted in chapters 5 and 6 of this thesis. Part of the work and materials in this chapter have been published in Alghamdi (2015), Alghamdi and Atwell (2017), and Alghamdi and Atwell (2018).

4.2 General extraction guidelines

In the development of AMWE lists the following main points were used as general guidelines in the AMWE extraction process:

In the preparation of reference data, consideration was given to the setting of the subsequent empirical AMWE extraction, particularly the evaluation methodology and the morphosyntactic selection patterns.

The AMWE extraction is based on the conceptual framework for AMWE described in chapter 3, which describes the practical definition and evaluation criteria of AMWEs.

Consideration was given to the rich morphology of SA described in section 3.2.1 and the designative properties of AMWEs described in section 3.4.

Because all the extracted AMWE items in this experiment will be manually evaluated,[41] the size of the extracted candidates needed for them to be feasible for manual annotation tasks will be considered. However, the amount of finally validated AMWE items should be significant enough to yield an acceptable performance estimation in the empirical evaluation tasks. Thus, frequency and linguistic filtering was implemented to reduce the extracted output to high-frequency and linguistically targeted items.

[41] More details about this type of evaluation are presented in section 2.3.1.

The final extracted lists should be classified into several datasets based on their POS patterns and the sequence length of AMWEs to facilitate their practical use.

4.3 The corpus source of the language data

The corpus used in this experiment was the ArTenTen corpus[42] (Arts et al., 2014), which contains more than 7.4 billion tokens. The corpus was automatically analysed using two different toolkits for SA morphological and linguistic disambiguation; the first was the Stanford Arabic Parser (SAP) (Manning et al., 2014) and the second was the MADAAMIRA toolkit (MA) (Pasha et al., 2014) for Arabic morphological and shallow syntactic analysis. This corpus was selected for several reasons, including its balance, representativeness, and size. The corpus was also considered to be representative of various written and spoken language genres. The corpus developers extracted their data from multiple online domains, which includes different semantic categories (e.g., science, politics, arts, and business).

In terms of data size, this is the most extensive and well-balanced SA corpus available for general purposes in corpus linguistics and NLP research. In corpus linguistic literature, several research studies emphasise the effect of corpus size in the overall improvement of language representation and the output quality of corpus-based and NLP experiments (e.g., Biber et al., 1999; Hunston, 2002; Lee and Cantos, 2002).

This does not mean this is the ideal corpus to work on, but within the constraints of the project it is the best practical and available large SA corpus. Finding a completely balanced and representative corpus remains difficult as McEnery and Hardie (2011 p. 10) explain that 'Balance, representativeness and comparability are ideals which corpus builders strive for but rarely, if ever, attain'.

The language data of the corpus was compiled from various web domains by the SpiderLing[43] Tool for web scribing. Table 4.1 provides essential information about the ArTenTen.

---

[42] The ArTenTen corpora can be accessed through the Sketch Engine website: https://www.sketchengine.co.uk.

[43] This tool is available through the following link: http://nlp.fi.muni.cz/trac/spiderling.

Table 4.1: Basic information about the ArTenTen corpus.

Data statistics	Number
Tokens	7,4 Billion
Words	5,7 Million
Sentences	177 Million
Documents	11.5 Million
Data size	58.0 GB

To the best of the researcher's knowledge, this is the largest available SA corpus of an acceptable quality and with detailed information about the corpus preparation and compiling processes. Most available SA corpora are limited in their size or the scope of SA representations. When it comes to corpus linguistics, these two criteria for corpus construction are considered the core elements in any corpus evaluation task (McEnery and Gabrielatos, 2008; Corpas, Pastor and Seghiri, 2010).

The ArTenTen corpus represents different SA domains and was divided into 28 sub-corpora according to the most common domains targeted by the web crawler during the corpus compiling process. The crawler tool used more than 116k domains to ensure comprehensive representations of SA; these domains were mainly from Arabic-speaking countries but also included several other countries with a large volume of SA websites. Table 4.2 shows the top 20 domains along with their percentages in the corpus.

Table 4.2: Top domains in the ArTenTen corpus.

Top domain	Percentage	Top domain	Percentage
.com	54.45	.cn	0.41
.net	20.86	.jo	0.4
.org	1.55	.sd	0.38
.info	1.41	.ma	0.35
.ps	0.76	.lb	0.3
.sa	0.61	.il	0.28
.sy	0.76	.biz	0.26
.eg	0.61	.ws	0.26
.ae	0.6	.ir	0.25
.cc	0.43	Other	4.03
.uk	0.41		

Before describing the methodology adopted in this study, a brief illustration will be given of the core components of SAP and MA toolkits implemented as part of the MWE extraction model. This is an essential step in understanding the outputs of linguistic analysis involved in the extraction process.

4.4 Automatic SA linguistic analysis toolkits

The ideal solution when creating gold standard evaluation LRs is to implement the MWE extraction model on a manually annotated corpus to avoid the possible errors usually associated with automatic linguistic tools. However, given the corpus size and the constraints of the project, this ideal situation is beyond the scope of the project for several practical reasons.[44] Thus, in the development of current AMWE datasets, most linguistic components in the discovery model were automatically implemented using two ANLP toolkits, SAP and MA, which are described briefly in sections 4.4.1 and 4.4.2.

[44] Several reasons justify the use of automatic linguistic analysis methods such as time limitations and labour-intensive work which require a dedicated expert team with sufficient funds. The reliance on automated linguistic toolkits are standard practice in NLP literature (e.g., Moirón, 2005b; Pecina, 2008; Seretan, 2011; Ramisch, 2012).

4.4.1 Stanford Arabic Parser (SAP)

This tool is part of the Stanford Core-NLP system (Manning et al., 2014), which is one of the most popular toolkits used in NLP research. The Stanford toolkits were developed initially for English NLP research and the toolkit developers later provided partial support for several other languages, including SA. The Arabic version supports with various quality the following NLP tasks:

Tokenisation and segmentation.

Part of speech tagging.

Sentence splitting.

Constituency parsing.

The SAP constitutes a linguistic pipeline that includes most core NLP tasks starting from text preparation, normalisation, and tokenisation to more complex and advanced functions such as syntactic parsing, semantic annotation, and conference resolution. However, the focus in this section is on the basic linguistic tasks applied by SAP in the extraction model, specifically SA tokenisation and POS tagging. The tokenisation of Arabic in SAP is based on the guidelines of the Penn Arabic Treebank annotation (PAT) (Maamouri and Bies, 2004). The PAT tokenisation is primarily based on the results of the morphological analyses generated by the Buckwalter Arabic morphological analyser (BAMA) (Buckwalter, 2004). Table 4.3 shows the POS tag set used by SAP[45]

Table 4.3: Basic POS notation of SAP.

Part-of-speech	Labels	Examples	
noun	(DT)?NN.*	مسجد	masjid
verb	VB.*	يذهب	yaḏhab
adjective	(DT)?JJ.*	جميل	jamīl
adverb	W?RB	بين	bayn
conjunction	CC	و/ف	wa/fa

[45] The complete notation and tagset are provided in Appendix C. For other useful information on SAP see: https://nlp.stanford.edu/software/parser-arabic-faq.shtml#d

preposition	IN	عن/ إلى	ʿan / ʾilā
pronoun	PRP.?	هو / أنت	huwa / ʾant
cardinal number	CD	الأول	alʾawwal

This tool adopted a tagset that classifies the words into eight core tags which include seven tags for the primary POS in SA, and the cardinal number tag which represents all numerical words in SA. Each one of these seven tags contains several sub-classifications that cover the principal morphological analysis. This increases the total number of tags used in SAP to 32 tags, as can be seen in Appendix C of this thesis.

Another vital point to consider at the tokenisation level of linguistic analysis by SAP is its treatment of cliticisation in SA. In this regard, the tool mainly separates clitics that play a role in the syntactic structure of the sentence. Thus, any clitics considered the subject or object in the sentence such as several types of pronouns should be separated from their attached words: clitics that do not affect the syntactic structure of the sentence, like the determiners (e.g., ال), remain attached to their words. Table 4.4 provides examples of the tokenisation variation of clitics in SAP according to their influence on sentence syntactic structures.

Table 4.4: Different clitics' tokenisation of SAP.

Clitics	POS	Examples	Tokenisation	Separation Mode
ه	Pronoun	جدارته	جدارت | ه	Yes
ب	Proposition	بالقوة	ب | القوة	Yes
و	Conjunction	ويبدو	و | يبدو	Yes
ال	Determiner	السلام	السلام	No

4.4.2 MADAMIRA Arabic morphological analyser (MA)

MA is another toolkit used for morphological disambiguation and linguistic analysis for ANLP tasks. Figure 4.1 illustrates the MADAMIRA[46] system Architecture.

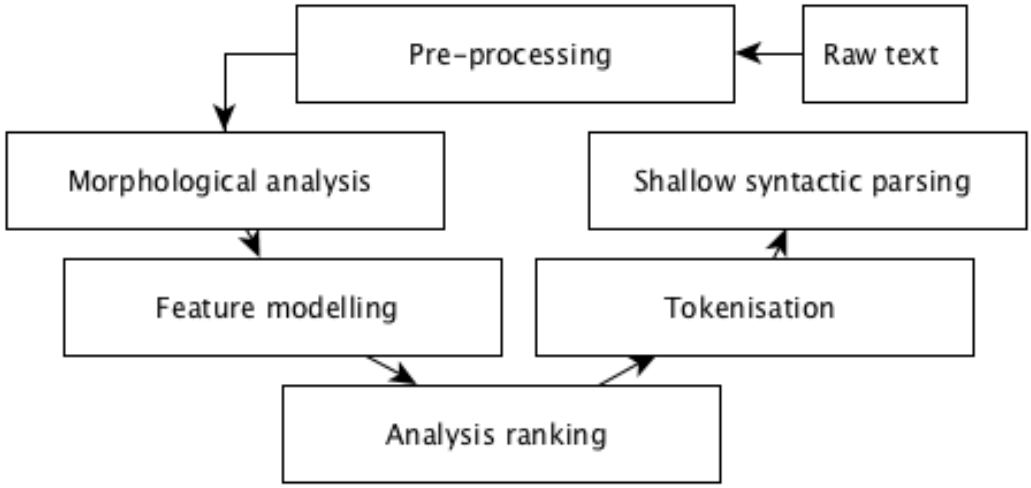


Figure 4.1: An overview of MA architecture (Pasha et al., 2014, p. 1095p.1095).

As shown, the toolkit pipeline of linguistic processing consists of seven phases which include the core morphological and syntactic tasks in SA, starting from cleaning and preparing the input data to shallow syntactic parsing and named entity recognition. Like the previous section, the focus is on the tokenisation and POS tagging part of MA. The POS tagset used by this tool constitutes 15 main tags, and several tags include several subcategories that represent in detail different types of morphological analysis in SA. Table 4.5 shows the core tags of MA along with their various subcategories.

[46] The MADAMIRA toolkit is publicly available and can be downloaded at http://innovation. columbia.edu/technologies/cu14012_arabic- language-disambiguation-for-natural- language-processing-applications.

Table 4.5: POS tag set for MADAMIRA (Al-Badrashiny et al., 2014).

POS	Labels	POS	Labels
Nouns	Noun	Foreign/Latin	latin
Number Words	noun_num	Abbreviations	abbrev
	noun_quant		
Proper Nouns	noun_prop	Punctuation	punc
Adjectives	Adj	Conjunctions	Conj
	adv_interrog		conj_sub
	adv_rel		
Adverbs	adv	Interjections	interj
	adv_interrog		
	adv_rel		
Pronouns	pron	Digital Numbers	digit
	pron_dem		
	pron_exclam		
	pron_interrog		
	pron_rel		
Verbs	Verb	Particles	part
	verb_pseudo		part_dem
Prepositions	prep		part_det
			part_focus
			part_fut
			part_interrog
			part_neg
			part_restrict
			part_verb
			part_voc

MA is an essential toolkit for any ANLP task because it supports the accomplishment of linguistic tasks in SA with adequate quality output. The following are the main processing tasks that can be conducted with MA:

Lemmatisation: determining the lemma

Diacritisation: determining the fully diacritised form

Glossing: determining the English glossary entry

Part-of-speech Tagging: determining the part-of-speech

Morphological Analysis: identifying every possible morphological interpretation of input words.

Full Morphological Disambiguation: determining a complete or partial set of morphological features (either the most likely feature values for each word given its context, or a ranked list of all possible analyses for each word).

Stemming: the reduction of each word to its morphological stem

Tokenization: segmentation of clitics with attendant spelling adjustments according to form.

A variety of schemes: the tokenisation scheme specifies the tokenisation separation rules and the output format (Pasha et al., 2014).

Due to their high precision and stable computational performance, SAP and MA are the most commonly used morphological toolkits in the ANLP research community. These toolkits are considered state-of-the-art in automatic linguistic analysis tasks, although recent experiments on a neural-based morphological system based on deep learning algorithms suggest a bright future for the improvement of morphological disambiguation toolkits in ANLP (e.g., Zalmout and Habash, 2017). Hence, the SAP and MA will be used in multiple phases of the AMWE extraction experiments and in the task of building a lexical model for the LR developed in this thesis.

4.5 Methodology: A Hybrid model for AMWE extraction

This model for extracting AMWE reference data combines statistical and linguistic components; hence, mixtures of several processing tasks were applied to retrieve AMWE items from large SA corpus. The primary objective of the extraction experiments was to produce AMWE datasets that will be used as the essential part of AMWEL and should also be beneficial as an evaluation LR in the following automatic extraction tasks. In this study, frequency data was used as one of the prime indicators for the usefulness of extracted candidates, following research on MWE which found this criterion to be an essential part of extraction models (e.g., Shin and Nation, 2008; Seretan, 2011; Ramisch, 2015a; Pecina, 2009).

The model consists of three core phases that result in the development of several reference datasets; in each stage, the extracted candidates undergo different sorts of

analysis until the final refined list of AMWEs is achieved. This model combined several extraction techniques following best practice in the literature within the constraints of the support available for computational processing of SA, as described in section 3.2. In the following subsections, several important issues related to the research methodology, along with a brief description of each AMWE extraction phase, will be presented.

4.5.1 Stages in constructing the AMWE reference datasets

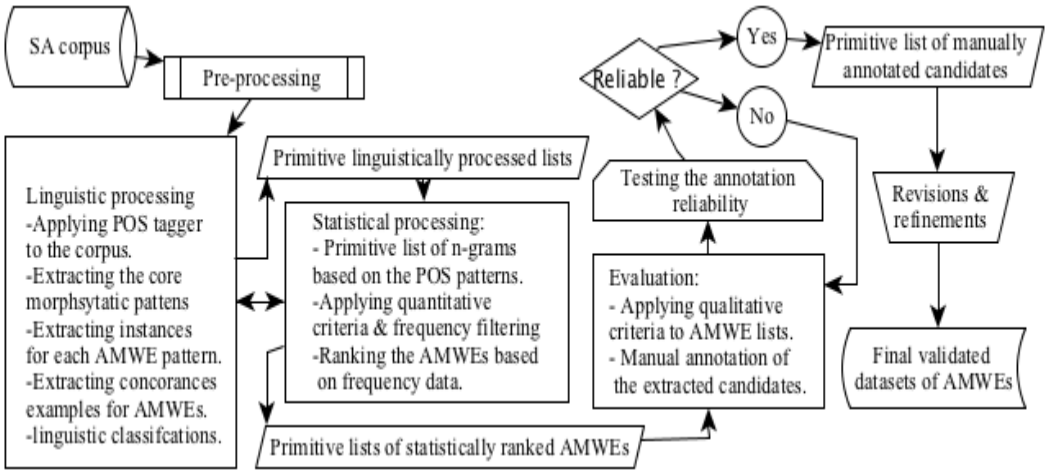


Figure 4.2: Diagram of the AMWE hybrid extraction model for reference datasets.

The AMWE extraction discovery model was implemented in a series of steps which consisted of three main stages, as shown in Figure 4.2: linguistic, statistical and evaluation phases. Each one of these stages consisted of several processing tasks which aimed to enhance the final extraction output as described in the following subsections.

4.5.1.1 Linguistic processing

Before describing the core linguistic components, it is important to note that, in the extraction model, there is no strict sequential order between linguistic and statistical processing tasks which is what is meant by the bidirectional arrow between these two main phases of processing. Thus, statistical methods in this model were occasionally used in the linguistic stage and vice versa.

The linguistic processing includes several tasks applied to enhance the extraction output based on linguistic disambiguation tasks such as text normalisation, tokenisation, lemmatisation, and POS tagging which were conducted using SAP and MA linguistic toolkits. Furthermore, this stage involves the extraction of multiple types of morphosyntactic structures in AMWEs based on the linguistic information. The retrieved AMWE patterns were then classified into several categories and sorted by their type and frequency data. Finally, the linguistically analysed data was saved in multiple tables for further statistical processing and candidate filtering tasks.

4.5.1.2 Statistical processing

This stage included several statistical components which primarily aimed to generate frequency and probabilistic data based on the linguistic information obtained from the previous extraction phase. This consisted of the use of n-gram models to extract a frequency-based list of morphosyntactic patterns and the extraction of AMWE instances based on predetermined AMWE selection structures. The extracted candidates at this stage were ranked in descending order according to their frequency. Moreover, this phase includes the implementation of frequency filtering of the extracted data to reduce the number of retrieved items so that the final lists could be manually annotated in the validation stage.

4.5.1.3 Evaluation and annotation

The qualitative evaluation phase aimed to manually classify the extracted candidates generated from the previous processing phases to true or false AMWE. This was based on detailed annotation guidelines founded on the adopted concept and criteria for AMWE (see chapter 3). Thus, any candidate which met at least one of the pre-determined qualitative criteria was included in the final list at this stage. Furthermore, to ensure the reliability of this manual annotation task, the inter-annotator agreement was measured and several annotation exercises conducted to test the reliability of the annotation. Finally, the validated AMWEs were linguistically classified into several datasets based on their morphosyntactic structures. Section 4.6 reports the experimental procedures in detail.

4.5.2 The extraction of discontinuous and nested AMWE candidates

Regular expressions and wildcard methods utilised in the extraction model allow the discovery of flexible AMWE sequences which include slots or gaps within the AMWE core lexemes. For instance, several plausible scenarios of the discontinuous V_N AMWE (ضَيَّقَ الخِنَاق, *ḍayyaq alẖināq*, tighten the noose), as shown in Table 4.6, were extracted by the following patterns:

([pos="?NN.*"][pos= ".*"][pos= ".*"][pos=".*"][pos= ".*"][pos="?NN.* "]).

Table 4.6: Examples of the possible slot within the AMWE *ḍayyaq alẖināq*.

Second lexeme	Examples of intervening sequences with their POS tags		First lexeme
	N-A-A	الحصار الاقتصادي الجائر	
	P-PRO-N	علي هم الشيوعيون	
	P-N-PRO	على نفس ه	
الخِنَاق	N-P-PRO	الخوف علي هم	ضَيَّقَ
	P-PRO-N-N	عليها رجال المباحث	
	N-N-A	فريق مانشستر يونايتد	
	N-N	وكيل النيابة	

The use of these techniques was limited to the discovery of sequences within a slot of 1 to 4 intervening lexical items. Thus, other discontinuous AMWE candidates with a more extended intervening slot were excluded because, based on corpus-based data, there were no AMWE candidates of interest with more than a four-word gap.

4.6 AMWE extraction Experiment

The extraction model implemented in this experiment consisted of three core stages which included several statistical and linguistic tasks applied to the corpus to arrive at representative AMWE datasets that cover various syntactic structures and semantic domains. The following subsections report the procedures implemented in the process of building well-validated reference datasets for AMWEs. A brief description will also be presented of the pre-processing phases and the automatic linguistic analysis conducted repeatedly in the extraction process. Furthermore, the data sources used for selecting the morphosyntactic extraction patterns will be explained and the computational treatment of discontinuous AMWE candidates highlighted.

4.6.1 Pre-processing phase

Normalisation is an essential task in the computational processing of SA text because the language script has several distinctive properties that might result in noisy data. Traditional normalisation tasks were conducted on the corpus to enhance the corpus quality, as described in section 3.2.1.1. However, in the experiments, function word types that could yield noisy data were retained as excluding functional classes would mean the discovery model misses an enormous number of valuable AMWE candidates. This choice was based on corpus-based evidence from a large SA corpus and by research in the literature that emphasises the importance of these types of words in MWE extraction tasks (e.g., Kato et al., 2013; van der Wouden, 2001).

At this stage, a blacklist of obsolete Arabic words developed by Attia et al. (2011) was also applied. The list contains around 8,400 words that are no longer used in contemporary SA text. Thus, the types of words in this list that are considered noisy data will not be part of any AMWE candidates of interest. Table 4.7 presents examples of obsolete lexical items from the list.

Table 4.7: Examples of obsolete Arabic words (Attia et al., 2011).

Transliteration	Words	POS	Translation
ʾarḫun	أَرْخُن	noun	archon
ʾarāḫina	أَراخِنَة	noun	notables
ʾarḫamīd	أَرْخَمِيد	noun_prop	Archimedes
ʾarḫamīdī	أَرْخَمِيدِيّ	adj	Archimedean
ʾarḫībūf	أَرْخِيبُوف	noun_prop	Arkhipov; Archipov
ʾirdabba	إِرْدَبَّة	NapAt	cesspool
ʾarduwāz	أَرْدُواز	noun	slate; board

The corpus was also cleaned of duplicated texts, misspelt words, and other types of noisy data that usually accompany the texts scribed from web-pages (e.g., text related to copyright, navigation panels, privacy notices, and commercial advertisements).

4.6.2 Automatic morphological analysis and POS annotation

In the following subsections, a brief description will be provided of the primary linguistic tasks implemented in the experiments based on the use of the automatic linguistic tool (SAP).

4.6.2.1 Linguistic processing with SAP

At this stage, the corpus was morphologically analysed and POS annotated using the SAP toolkit. The morphological analysis was based on the use of the BAMA analyser[47] (Buckwalter, 2004) which includes a comprehensive Arabic morphological lexicon consisting primarily of three Arabic-English lexicon files and three morphological compatibility tables. The core linguistic information of BAMA and several sample entries from the lexicon are presented in Tables 4.8 and 4.9.

Table 4.8: Basic information about the BAMA analyser.

Type	Number
Prefixes	299 entries
Suffixes	618 entries
Stems	82158 entries
Lemmas	38600 lemmas
Prefix-stem combination tables	1648 entries
Stem-suffix combination tables	1285 entries
Prefix-suffix combination tables	598 entries

Table 4.9: Sample entries from the BAMA morphological lexicons.

Examples of Arabic prefixes and their concatenations			
b	bi	NPref-Bi	by;with <pos>bi/PREP</pos>
k	ka	NPref-Bi	like;such as <pos>ka/PREP</pos>
Al	Al	NPref-Al	the <pos>Al/DET</pos>
Examples of Arabic suffixes and their concatenations			
p	ap	NSuff-ap	[fem.sg.] <pos>ap/NSUFF_FEM_SG</pos>
tynA	atayonA	NSuff-tay	two [acc.] + our <pos>atayo/NSUFF_FEM_DU_ACC_POSS+nA/POSS_PRON_1P</pos>
tykmA	atayokumA	NSuff-tay	two [acc.] + your [du.]

[47] This can be downloaded from the linguistic data consortium (LDC) at: https://catalog.ldc.upenn.edu/ldc2004l02

Examples of Arabic stems			
ktb	katab	PV	write
ktb	kotub	IV	write
ktb	kutib	PV_Pass	be written;be fated;be destined

The BAMA morphological lexicon used in the development of many ANLP toolkits (e.g., Marton et al., 2013; Pasha, 2014). BAMA is frequently used because it represents most morphological features of SA and provides the appropriate representational information to facilitate the integration process in different ANLP tasks. The following subsections illustrate the main components of linguistic analysis conducted in this study, which includes tokenisation, lemmatisation, and POS tagging.

4.6.2.1.1 Tokenisation

The tokenisation of SAP is mainly based on the morphological analysis provided by BAMA which was also used in APT (Maamouri and Bies, 2004). Regarding the treatment of cliticisation in Arabic, SAP primarily treats most clitics that affect the syntactic structure of the sentence as separated tokens. For instance, based on the SAP tokenisation analysis, the object and subject pronouns are cliticised in the verbal phrase, فهمتها, fahimtuhā, *I understood it*.

Other clitics that do not affect the syntactic structure remained attached to their adjacent words, as was the case for the determiner ال, al, *the* in SA which is considered by the tool to be part of the attached token. Furthermore, inflectional and derivational forms were not separated off by the default tokenisation configurations. Figure 4.3 presents an example from the APT where the object pronoun was همseparated from the preposition من because of the syntactic function of هم as an object in this phrase.

min-fromمن(PP

hum-them_[masc.pl.]))هم(NP

min + hum

from + they

from them

Figure 4.3: Prepositional phrase with a cliticisd object pronoun *hum* which splits apart from the preposition *min*, (Maamouri et al., 2009).

4.6.2.1.2 Lemmatisation

Lemmatisation is an essential task to implement in the extraction model to produce more precise statistical information about the generated AMWE candidates, especially in morphologically rich languages such as SA. Thus, in the extraction, the analysis was based in most cases on the lemma of words which group several related forms to their core lexemes. However, in several cases, this task might result in excluding the extraction of useful AMWE candidates because of the low-quality lemmatisation output; the lemmatisation task was therefore occasionally applied after the extraction task as part of the filtering processes for candidate lists. However, the computational task for SA text lemmatisation still faces many limitations and problems in the analysis of the final output due to the complex and rich morphological system. Therefore, because the SAP toolkit does not provide support for lemmatising Arabic text, additional toolkits specifically available for Arabic text segmentation and lemmatisation were used (Smrž, 2007; Darwish and Mubarak, 2016; Pasha et al., 2014).[48]These toolkits were used in the extraction and filtering processes to achieve the best possible outputs in this study. Table 4.10 presents an example of the AMWE N_N شحذ الهمم, *šaḥaḏ alhimm, sustain the momentum* with related forms found in the corpus. More details about this task are given in section 2.2.1.2.

Table 4.10: An example of inflectional forms related to the core lexemes of AMWE.

POS	Noun	Noun
Lexeme	هم	شحذ
Inflectional forms	همم	يشحذ
	هممنا	تشحذ
	هممكم	اشحذوا
	الهمم	يشحذون
	الهمة	تشحذ
	همتهم	شحذوا
	همتي	شحذت

[48] All these toolkits are open-source projects. For more details and downloads, visit the following links: https://github.com/otakar-smrz/elixir-fm.,http://qatsdemo.cloudapp.net/farasa, http://nlp.ldeo.columbia.edu/madamira.

4.6.2.1.3 POS tagging

This is the most critical linguistic task and one that plays a significant role in the improvement of the AMWE extraction model. By annotating the raw corpus text with POS tagging, invaluable information can be extracted about various AMWE patterns. Figure 4.5 illustrates the POS distributions of all words in the corpus after the automatic morphological analysis and annotation were implemented using the SAP toolkit. The POS data shows that nouns in their various forms are the dominant POS category with more than 3.5 billion tokens followed by a verb with less than 1 billion tokens. Conjunction, preposition and pronoun constitute a similar size of approximately 500 million tokens while adjective and cardinal number tags annotate more than 250 million tokens. The lowest tagged word class was the adverb with less than 100 million tokens in the corpus.

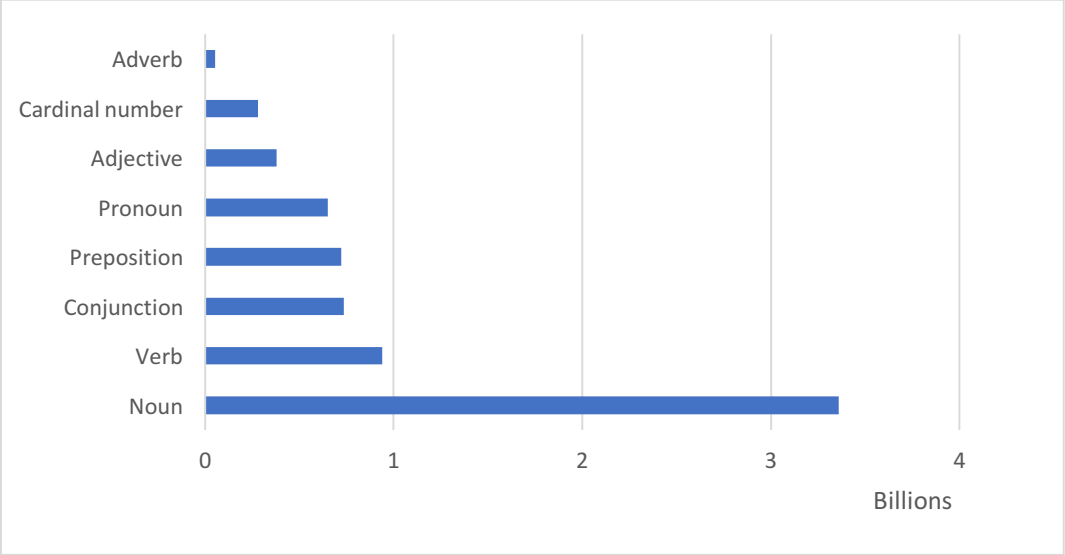


Figure 4.4: POS distribution after the automatic morphological analysis using SAP.

These are the main POS classes used by the SAP. There were 32 morphological tags in total which included various subcategories of the core POS classifications. Figure 4.5 shows the disruptions of the most frequent POS tags based on their average reduced frequency (ARF).

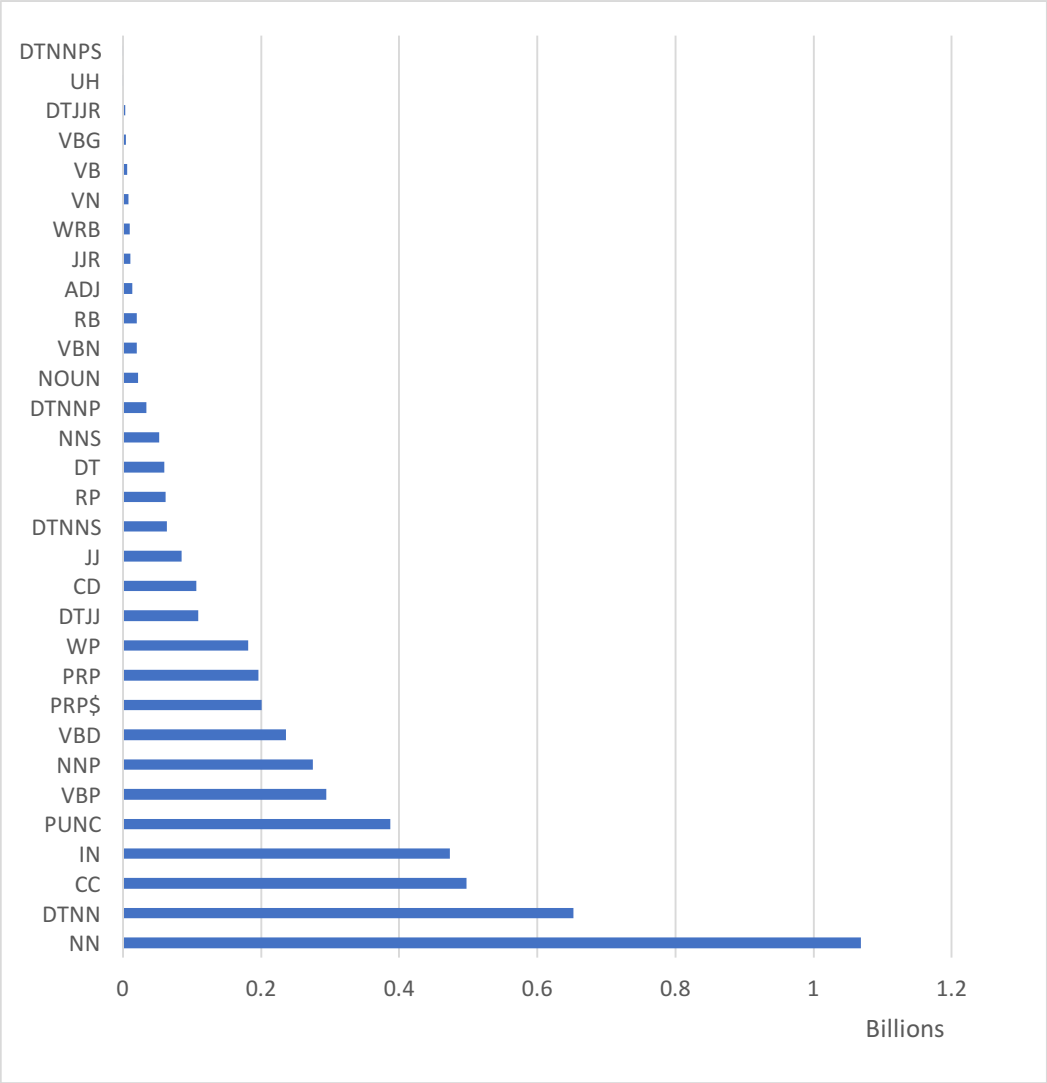


Figure 4.5: The disruptions of high frequency POS tags in the corpus based on ARF.

As expected, nouns and verbs constitute the vast majority of POS classes in the data based on the output of POS tagging conducted in this phase; thus, it is estimated that a large number of nominal and verbal expressions will require more attention in the pattern extraction process in this study. At the end of this processing phase, the corpus was linguistically analysed by applying the core tasks which included tokenisation lemmatisation and POS annotation. These linguistic processing tasks are essential in preparing the data for the next level of the extraction model, which relates the discovery of various morphosyntactic patterns and their AMWE instances from the annotated corpus using statistical and other corpus search techniques.

It is worth noting that the use of automatic linguistic analysis tools usually yields several types of errors in processing, as mentioned in earlier tasks. These generated errors affect the performance of the extraction model by adding unwanted items or removing useful MWE candidates. Hence, in the experiments, an analysis of the types of errors found in the data will be presented where appropriate to eliminate their impact on the final extraction output. Evert and Kermes (2003) state that most automatic analysis errors can be found in the extraction of low-frequency items. Therefore, it is important to be aware of these potential errors, especially when dealing with less frequent candidates.

4.6.3 Selecting the AMWE extraction patterns

The large volume of AMWEs found in the corpus and the limited scale of the current study requires the imposition of several morphosyntactic constraints in the AMWE patterns extraction task. However, rather than merely using intuition in deciding the best selection patterns to use for extracting invaluable AMWE candidates, the choice was based on more reliable sources (MWE literature, and linguistic and statistical information in the corpus), as will be illustrated in the following subsections.

4.6.3.1 MWE literature

Based on the assumption that AMWE constructions common in other languages might also be frequent and yield interesting AMWEs in SA, the core extraction patterns found in MWE research were reviewed and their usefulness and feasibility in AMWE extraction process examined. However, no consensus can be found in the literature on specific morphosyntactic patterns used in most computational extraction experiments. Thus, the selection method was for extraction patterns mostly affected by the distinctive linguistic properties and the statistical information of the targeted language. Table 4.11 shows several examples of extraction sequences used in several MWE studies. More details about related studies on MWE patterns are discussed in sections 3.4, 3.5 and 2.2.1 of this thesis.

Table 4.11: Examples of MWE extraction patterns used in the literature.

Research	Examples of MWE Patterns used
(Smadja, 1993)	[A-N] [N-N] [S-V] [V-O] [V-P] [V-Adv]
(Basili et al., 1994)	[A-N] [N-N] [N-P-N] [S-V] [V-P]
(Benson et al., 1997)	[A-N] [N-P-N] [S-V] [V-O] [V-P-N] [Adv-A] [V-Adv]
(Lin, 1998)	[A-N] [N-N] [S-V] [V-O]
(Kilgarriff and Tugwell, 2001)	[A-N] [N-N] [N-P-N] [S-V] [V-O] [V-P]
(Goldman et al., 2001)	[A-N] [N-N] [N-P-N] [S-V] [V-O] [V-P] [V-P-N]
(Korkontzelos, 2010)	[N-N] [A-N-N] [A|N[N-P]A|N] [A|N-N]
(Seretan, 2011)	[N-A], [A-N], [N-N], [N-V], [V-N]
(Ramisch, 2015)	[Adv-A], [A-N], [V-P], [V-N], [N-N]

The divergence in the extraction patterns used in these studies might also be due to the insufficient corpus-based research on MWE which provides evidence of MWE linguistic behaviour within languages. Furthermore, these variations also reflect the widespread use and heterogeneous nature of MWE phenomenon which are explained by the various morphosyntactic constructs.

In finding the most predictive selection patterns, the focus lay specifically on reviewing previous AMWE research that presents valuable information on the most frequently used selection patterns in SA. Table 4.12 presents examples of the patterns used in AMWE research. However, a similarly diverse finding regarding the used selection patterns was once again observed.

Table 4.12: Examples of extraction patterns used for AMWEs in the literature.

Research	MWE Patterns
(Elewa, 2004)	[V-N] [N-N][N-C-N]
(Cardey et al., 2006)	[V-N] [N-N]
(Boulaknadel, et al. 2008)	[N-N], [N-A],[N-P-N]
(Attia, 2008)	[N-A], [N-N], [N-N-A], [P-N-N], [P-N]
(Bounhas and Slimani, 2009)	[N-N], [N-A], [N-P-N],[N-P-N-N]
(Attia et al., 2010)	[N-N], [N-A], [N-P-N],[N-C-N]
(Saif et al., 2011)	[N-N], [N-A], [N-V], [V-Adv], [A-Adv] [A-N]
(Abdou, 2011)	[V-N-N], [V-N-P-N], [V-C-V],[N-C-N], [N-N-N], [A-N P-N], [P-N-A]

Furthermore, corpus-based and traditional Arabic linguistic studies were used to present a detailed analysis of multiple types of basic AMWEs and their various morphosyntactic structures (e.g., Abdou, 2011; Elewa, 2004).

4.6.3.2 Linguistic and statistical information from the corpus.

The most crucial source for AMWE patterns is the linguistic and statistical information obtained from empirical observations in the corpus-based analysis conducted through AMWE preliminary experiments. These yielded substantial evidence and statistical data about the actual use of AMWEs and can be used as an indicator of the most productive AMWE patterns.

In the process of selecting the morphosyntactic extraction patterns, all the sources mentioned above were combined to produce the best possible selection patterns for discovering AMWE candidates. The selection was conducted as an iterating process, as illustrated in Figure 4.6, and includes three preparing the possible selection patterns, and then using them in several trial extraction experiments. Finally, based on the output quality of the trail extractions, the patterns were either added to the extraction model or the selection process was restarted to find AMWE patterns that were more predictive.

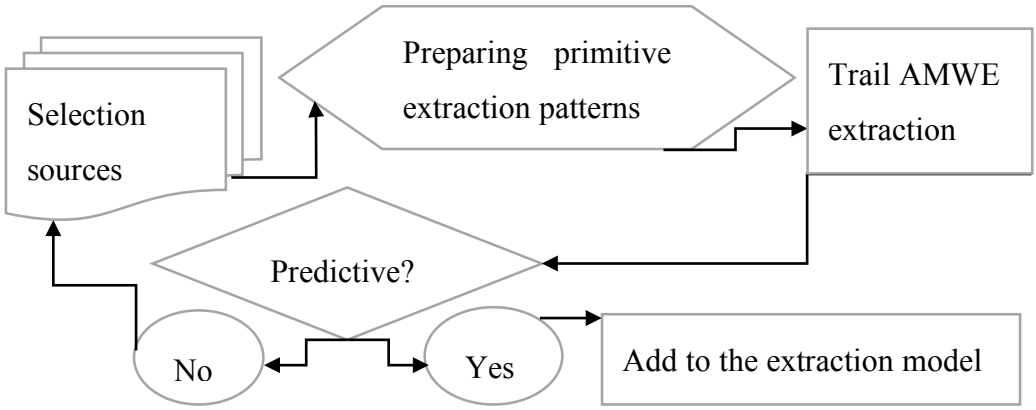


Figure 4.6: The iterating process for selecting AMWE extraction patterns.

4.6.4 Statistical processing

In this phase of processing, several statistical tasks were conducted to extract multiple lists of potential AMWE patterns and instances from the linguistically annotated corpus. Initially, the n-grams model was used to retrieve several lists of POS patterns which ranged from 2 to 6 n-grams. The retrieved lists of morphosyntactic patterns were then saved in multiple files and classified based on their frequency and linguistic information. Following the patterns selection process described in section 4.6.3, a set of the morphosyntactic patterns was used to extract AMWE instances in multiple trial experiments to explore the productivity of various common AMWE patterns that can be utilised in the extraction model. Figure 4.7 presents examples of the most frequent POS patterns extracted based on the SAP tagset used in the linguistic annotation of the corpus. This shows only five examples of each n-gram category from the extracted patterns which represents a small sample of the data. For a more extensive list of extracted patterns along with statistical data, see Appendix E.

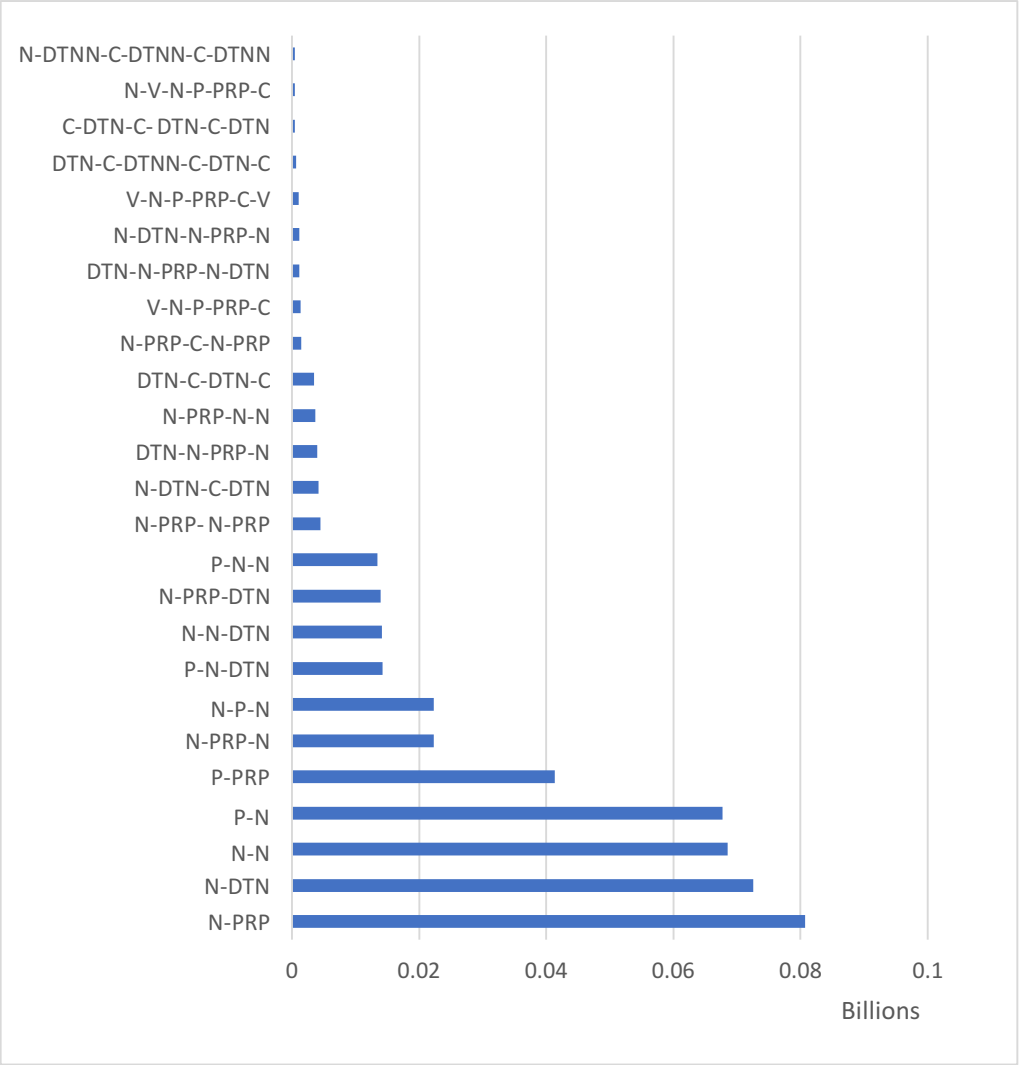


Figure 4.7: The five most frequent POS patterns (2 to 6 n-grams) automatically extracted from the corpus based on morphological analysis by SAP.

The data shows that nominal and prepositional phrases are dominant among the most frequent POS patterns in the corpus. As expected, the shorter the expressions, the more frequently they were found in the corpus and vice versa. These findings regarding MWE POS patterns are in line with the findings of several MWE research studies, particularly AMWE extraction experiments (section 2.4) on related MWE LRs.

Following the linguistic and statistical analysis of the potential extraction patterns and the use of several selection patterns from other sources described in section 4.6.3, the final list of extraction patterns was produced and were used in the final extraction model to build a reference dataset of AMWEs. The morphosyntactic patterns were then classified into three categories based on the adopted typology of AMWEs

illustrated in section 3.5.6. Table 4.13 shows the patterns used in the extraction model that represent various morphosyntactic structures.

Table 4.13: The extraction patterns with candidates from corpus-based instances.

Classes	No. of tokens	AMWE patterns
Nominal	2	[N-N] [N-A]
	3	[N-C-N] [N-P-N]
	4	[N-N-P-N] [N-N-N-A]
	5	[N-N-P-N-N][N-N-A-P-N]
Verbal	2	[V-N] [V-P]
	3	[V-P-N][V-N-N]
	4	[V-P-N-N][V-N-N-N]
	5	[V-N-N-C-N] [V-N-N-P-N]
Prepositional	3	[P-N-N] [P-N-A]
	4	[P-N-N-N] [P-N-N-A]
	5	[P-N-N-C-N] [P-N-N-P-N]
	6	[P-N-N-P-N-N] [P-N-N-A-P-N]

Based on the conceptual framework for AMWE described in chapter 3, it is assumed that AMWEs can be found in various syntactic constructions in SA. Thus, in the extraction experiment, a wide range of syntactic structures will be considered to extend the coverage of AMWE constructs included in the final reference lists. The total number of extraction patterns used at this stage was 60 POS patterns, which were classified into nominal, verbal and prepositional expressions based on the prime typology of AMWEs.

Regarding the number of tokens included in the extraction, coverage was extended to phrases that consist of two to five core lexemes; however, this length restriction does not apply to discontinuous expressions that might have intervening words arranged from one to four tokens in the extraction of flexible AMWEs. Nevertheless, each pattern for these main constructions includes several variations that reflect the morphological variety of the forms of candidates. For instance, Table 4.14 presents a list of detailed patterns included in the core patterns N-N.

Table 4.14: Examples of multiple variations of AMWE [N-N] patterns.

Main patterns	Structural variations
N-N	NN DTNN
	NN NNP
	NNP NN
	NNP NNP
	NNP DTNN
	NNS NN

However, at this stage of AMWE extraction, there was no specific class for adjectival or adverbial expressions because, based on the pilot extraction experiments, a large number of phrases belonging to these types in the corpus could not be found. Thus, the extracted types of expression were included as a subclass under the prepositional expressions category.

4.6.5 Using extraction patterns to discover AMWE instances from the corpus

The selection of extraction patterns from various sources was based on their productivity in generating valid AMWEs. During this phase of processing, morphosyntactic extraction patterns were used to extract a list of instances for each pattern, which resulted in the generation of multiple large lists of AMWE candidates with a total of more than 60k items. The retrieved items represent a variety of lexical and semantic domains; however, this vast number of candidates renders manual evaluation a time-consuming and challenging task which, in this experiment, meant that statistical constraints had to be applied to limit the number of extracted AMWE instances, as will be illustrated in the candidate filtering phase in section 4.6.6. Table 4.15 presents examples of retrieved candidates that represent various selection structures. More examples along with statistical data for the AMWE candidates are provided in Appendix F.

Table 4.15: Examples of AMWE candidates extracted from the corpus.

AMWEClasses	Nu.	POS patterns	Instances	
Nominal	2	N-N	سقوط النظام	suqūṭu anniḍām
	3	N-P-N	عصفورين بحجر	ʿuṣfūrayni biḥajar
	4	N-N-C-N	اهل الحل و العقد	ʾahlu alḥalli wa alʿaqd
	5	NN-P-NN	الرجل المناسب في المكان المناسب	arrajulu almunāsibu fī almakāni almunāsib
Verbal	2	V-N	نصب الراية	naṣbu arrāyat
	3	V-N-Adv	ترى النور قريباً	tarā annūru qarīban
	4	V-Adv-P-N	تعمل جنبا إلى جنب	taʿmalu janban ʾilā janb
	5	V-N-Pro-P-N	تفتق ذهنه عن فكرة	tafattaqa ḏihnuhu ʿan fikra
Prepositional	3	P-N-N	على قدم المساواة	ʿalā qadami almusāwāti
	4	P-N-C-N	بجد واجتهاد	bijiddin wajtihād
	5	P-N-N-P-N	بعين الاحتقار والازدراء	biʿayni aliḥtiqāri wālizdirāʾ
	6	P-V-N-P-N-N	لنقدم عربونا على صدق النية	linuqaddima ʿurbūnan ʿalā ṣidqi anniya

For a list of high-frequency POS patterns, several tactics were used to retrieve AMWE candidates with gaps, as described in section 4.5.2. Table 4.16 displays a list of examples of discontinuous candidates extracted by multiple regular expressions.

Table 4.16: Sample from the retrieved flexible AMWE candidates.

POS pattern	Discontinuous candidates	
V-N-N-P	أعربت السيدة انا تيباجوكا عن	ʾaʿrabat assayyidatu anā taybājūkā ʿan
V-N-N-N	قدمت الملاعب السعودية عددا	qaddamat almalāʿibu assuʿūdyyatu ʿadadan
V-N-N-P-N	يستعد المنتخب التونسي لخوض	yastaʿiddu almuntaḫabu attūnisiyyu liḫawḍ
N-A-P-A-N	نهاية محتومة في آخر المطاف	nihāyatun maḥtūmatun fī ʾāḫiri almaṭāf
V-N-Adv-N-P-N-A	اعطي الفرصة تلو الفرصة للفريق الاخر	ʾaʿṭā alfurṣata tilwa alfurṣati lilfarīqi alʾāḫar
N-C-N-A-A	المشتقات والمنتجات النفطية المختلفة	almuštaqqāt walmuntajāt annifṭiyya almuḫtalifa
V-N-A-P-V-P-N-Pro	يكفي الشيخ فضلا ان يذكر عن وزيره	yakfī aššayḫ faḍlan an yaḍkur ʿan wazīrih
Pro-P-N-N-C-N	هم تحت خط البؤس والفقر	hum taḥta ḫaṭṭi albuʾs walfaqr
P-N-N-C-N-A-P-N	من مناطق الفقر وخاصة المدقع بالبلد	min manāṭiq alfaqr wa ḫāṣṣatan almudqiʿ bilbalad

As shown, the extraction patterns used for these items cover various types of intervening words in multiple places within the phrase, including the initial, middle, and final parts of the candidates. Interestingly, during the process of extracting AMWEs with gaps, in several cases the model discovered new, related AMWE candidates that were used to accompany the expressions extracted initially, such as

the two expressions in examples four and five, *aʿṭā alfarṣa* and *nahāya maḥtūma*. The outputs of this stage are lists of AMWE candidates based on multiple POS selection patterns. These, along with their linguistic and statistical information, then underwent various filtering tasks, as illustrated in the next phase of processing.

4.6.6 Candidate filtering

The filtering process is an essential step in refining the initially generated lists of candidates to control their size, exclude noisy data, and eliminate the number of false AMWE items. All tasks in this phase were executed automatically by using the multiple ANLP toolkits available. The fundamental filtering processes were implemented to prepare the datasets for the evaluation and manual annotation tasks as follows. Initially, all the items that contained spelling errors or inappropriate linguistic annotation in the extracted data were removed. The output of this refinement process was a list of 51,482 candidates which were ranked in descending order according to their normalised frequency in the corpus. The linguistic filtering includes the exclusion of 24 patterns from the extracted candidates because corpus-based exploration of multiple samples shows only a few valid AMWEs in the removed selection patterns. This process removed more than 13,743 instances from the retrieved lists. In another candidate filtering task, several open-source tools were automatically used to identify NEs and remove them from the extracted files (Schneider et al., 2013; Boudlal et al., 2010; Darwish and Mubarak, 2016). This resulted in the removal of more than 2479 NEs items from the retrieved lists.

Regarding statistical filtering, for each pattern a frequency threshold of various frequency scores was applied based on the number of components of candidates. This task resulted in the retention of 17,382 extracted AMWE candidates that were then evaluated and validated to construct the final refined list of AMWE that will be used in various evaluation and NLP applications, most notably as gold standard datasets for the subsequent AMWE extraction experiments reported in chapters 5 and 6. Furthermore, the filtering methods for the extracted candidates in this study have been used in several other studies and are a practical method for filtering out a significant amount of unwanted items in the retrieved data(e.g., Evert and Krenn, 2005; Evert

and Krenn, 2001; Smadja, 1993a; Pearce, 2002). Table 4.17 presents examples of the items removed from the outputs by the filtering processes.

Table 4.17: Sample of items removed by multiple filtering tasks.

Removed candidates		Filtering methods
ن عباد	a ʿabād	Erroneous and noisy items
الاحوط ل زوما	alāḥūṭ lu zawmā	
حيث تضع ك ل قدم	ḥayṯ taḍʿ ka la qadam	
الشعب ل ا ياكل الا ف ي رمضان	aššaʿb la aa yākul alā fa ya ramaḍān	
هنا يصف القرد ب عض	hunā yaṣif alqird ba ʿaḍ	
اللبنانية سوزان	allabnānya sūzān	Linguistic filtering
اخر غير	aḫar ġayr	
الى الله لكن	alā ʾallāh lākin	
و يقال ل هم	wa yuqāl la hum	
ف قال رسول الله صلى الله	fa qāl rasūl ʾallāh ṣallā ʾallāh	
و زن الراس الواحد من	wa zan arrās alwāḥid man	Statistical filtering
بعينك يرتكب ذنبا	biʿaynak yartakib ḏanbā	
ب عين الكرش	bi ʿayn alkarš	
كي اصرخ	kay aṣruḫ	
شخص معين او اشخاص	šaḫṣ muʿayyan aw ašḫāṣ	
قطاع غزة	qiṭāʿ ġazza	NEs
شمال افريقيا	šamāl afrīqiyā	
دولة قطر	dawlat qaṭar	
المملكة المتحدة	almamlaka almuttaḥida	
سلطنة عمان	salṭanat ʿumān	

4.7 Evaluation and annotation

As mentioned in section 2.3, several methods have been suggested in the literature for evaluating MWE extraction models and validating reference datasets (e.g., Evert and Krenn, 2001; Seretan, 2011; Luiz et al., 2011; Ramisch et al., 2012; Carpuat and Diab, 2010). However, there is no consensus regarding a specific approach that should be followed in the evaluation of various extraction output. Therefore. several factors should be considered in the selection of a particular method, most of which relate to the nature of the extraction task and the specific requirements of the targeted LRs and applications.

This diversity in preferences for a specific evaluation method does not imply there is no standard evaluation practice in this research area. Thus, the typical use of several evaluation methods can be observed, most of which are borrowed from information retrieval fields, including precision, recall, F-measure, and the mean average precision (MAP) scores. These are also used intensively in the evaluation of most NLP tasks. Furthermore, the uninterpolated average precision (UAP) is another method used in several MWE evaluation experiments (e.g., Seretan, 2011; Pecina, 2009; Moirón, 2005a). This is a combined set of precision measures that results in one evaluation score and reflects the precision of the extraction model and, indirectly, the recall score in the evaluation process. The test sets involved in the evaluation are usually based on random sampling from the extracted candidates or one or more n-best lists based on various extraction types. In the evaluation, as many extracted candidates as possible were used in the evaluation task within the constraints of the study because the ultimate aim of the experiment was to generate a reference list of AMWEs that can be used in the evaluation of further AMWE discovery studies and other NLP and language-related tasks.

However, given the factors mentioned above, manual classification and expert judgment was adopted as the evaluation method in this experiment, mainly due to a lack of well-validated AMWE evaluation datasets that can be used in parallel with manual annotation to accelerate the evaluation process. As illustrated in section 2.3, the available AMWE LRs are either not available as open source data or have limited coverage of the targeted AMWE types included in the AMWE frameworks described in section 3.3. This evaluation method has been used in several previous research studies and results in well-validated and high-quality MWE datasets (e.g., Seretan, 2011; Evert, 2004; Pecina, 2009).

In the annotation task, the annotators were asked to classify the extracted candidates into true or false MWEs based on the detailed selection guidelines described in sections 3.3.2 and 3.3.3. Further details about the annotation procedures are presented in section 4.7.1. To ensure the reliability of this task, common practice and recommendations regarding manual annotation were adhered to which includes writing detailed guidelines, descriptions of the tasks, and illustrating the AMWE concept and selection criteria which is a fundamental step in achieving reliable

agreement between annotators. This reliability testing process is illustrated in Figure 4.8 and shows that the annotation task should start with clear guidelines which explains the annotation objectives and the detailed procedures involved. Multiple pilot annotation tasks are then undertaken to measure the reliability and extent of inter-annotator agreement until a satisfactory level of agreement is achieved, after which the participants should be ready to start the main annotation process.

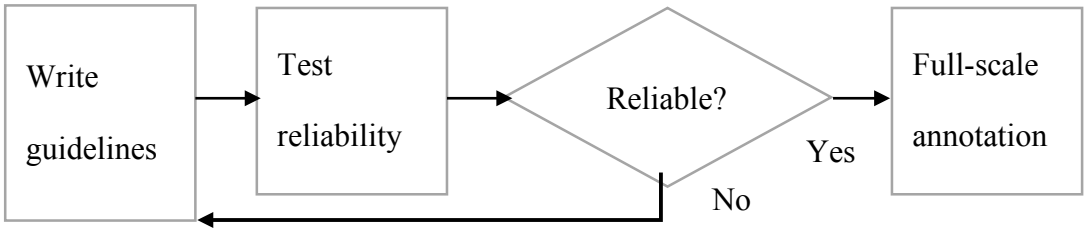


Figure 4.8: The process of reliability testing for manual annotation tasks (Ide and Pustejovsky, 2017, p. 299).

4.7.1 Annotation procedures and guideline

The manual annotation task in this experiment evaluated a sample from the AMWE extraction model's output to measure its performance and generate final validated datasets of AMWEs that can be used as reference data in the following extraction experiments and as part of the large-scale AMWE computational lexicon. The 17,382k AMWE candidates generated from previous processing phases underwent manual classification in this evaluation phase. Section 4.7.1.1 briefly describes the annotation guidelines provided to the coders who participated in this evaluation task, while sections 4.7.1.2 and section 4.7.1.2 report the annotation procedures and summarise the evaluation results for this extraction experiment.

4.7.1.1 Annotation guidelines

Based on the selected working definition and the linguistic properties of AMWEs described in sections 3.3 and 3.4, the annotation guidelines provided to the annotators who conducted manual evaluation of the extraction experiment will be briefly explained.

Most of the criteria used in this study have been adopted through an intensive analysis of previous English and Arabic research on manual extraction and classification of MWEs (e.g., Leech et al., 2001; Wray and Namba, 2003; Durrant, 2008; Shin, 2008; Wray, 2009; Ellis, 2010; Schmitt and Martinez, 2012; Ackermann and Chen, 2013). For instance, Wray and Namba (2003) proposed a set of eleven criteria that assist the researchers to use their intuitive judgment in the manual evaluation of MWE items. Martinez (2011) also outlined six core and auxiliary criteria to be used in the manual selection of MWEs. More in-depth criteria presented by Moirón (2005a) for finding the linguistic features of fixed MWEs is shown in Table 4.18, which summarises the linguistic properties of potentially fixed phrases. However, most of these features can be found in AMWE with few variations.

Table 4.18: Linguistic features for selecting fixed expressions (Moirón, 2005a, p. 48).

Lexeme level	
ordinary lexemes	high co-occurrence frequency
	peculiar meaning only present in a fixed expression
nonce words	only exist in the fixed expression
morphology	
inflectional	singular/plural morpheme in nouns
	diminutive
	if adjective gradable
	archaic forms (-e ending)
	case marking (determiners)
	tense inflection
derivational	prefix, compounding
Semantic	
denoting vs non-denoting lexemes	only literal meaning
	polysemous
	non-denoting

opacity	transparent
	semi-transparent
	semi-opaque
	fully opaque
conventionality	
compositionality	
decomposability	
Syntactic	
morpho-syntactic structure	regular structure
	synt. marked
	ill-formed
internal variation	modification, quantification, determiners
intervening adjuncts	between required constituents
agreement relations	with subject/object
syntactic versatility	topicalization, passive, . . .
open slots	words/phrases
non-homomorphism (syntax-semantics interface)	

These criteria, along with others suggested by previous research, were therefore considered when developing a set of criteria for this task. The main challenge in trying to establish a set of selection criteria for annotators concerned how to outline clear-cut criteria for selecting MWE. This was a hard task to achieve because of the complexity and heterogeneous nature of this linguistic phenomenon. Therefore, practical criteria were adopted that can be applied by several coders with an acceptable degree of inter-agreement reliability. Thus, the researcher set the following main criteria for annotating AMWE candidates.

In the annotation process for classifying AMWEs, any candidates that met at least one of these criteria should be considered a true AMWE that can be added to the final reference data. However, what all the requirements had in common was that they were established to help justify why it was believed the expressions chosen might pose some difficulty for any NLP task based on linguistic and semantic properties. The criteria for AMWE annotation were as follows:

Does the expression, or part of it, lack semantic transparency? This means that the meaning of the phrase is not derived from its component parts, such as ‘kick the bucket' which means to die, and in Arabic انتقل إلى رحمة الله *intaqal ’ilā raḥmat*

ʾallāh‘passed to the mercy of God’ which means مات, *māt* ‘die'. However, fully semantically transparent phrases are rare in language. Therefore, expressions with any degree of non-compositionality were taken into consideration in the identification of AMWEs.

Is the expression a morpheme equivalent unit? This criterion is concerned with the expression of ‘a holistically stored single lexical unit because its meaning and function map onto the form as it stands', p32). For instance, ‘in order to' and in Arabic على الرغم من, *ʿalā arraġm min ‘although’*.

Is the expression related to a specific situation or register? In every language, many expressions are firmly attached to particular occasions that are usually used to convey a precise meaning related to the situation, such as, ‘excuse me' and ‘happy birthday' and in Arabic شكرا لك, *šukran lak* ‘thank you’ and مع السلامة, *maʿa ssalāma ‘goodbye’*.

Does the expression exhibit an irregular grammatical structure? This includes phrases that are inconsistent with language rules, such as the expression ‘by and large’ in English. In Arabic, several fixed MWE sviolate the grammar roles as can be seen in the AMWE في حيص بيص *fī ḥayṣ bayṣ* ‘in confusion’, where حيص بيص has no case endings regardless of its context or position in the sentence.

Can the expression be paraphrased or translated into a single word? This criterion helps identify a MWE. In English, several studies have used a translated corpus to detect a different kind of MWE (Nerima et al., 2003, Smadja et al., 1996) by analysing their equivalent in other languages, for instance, the Arabic phrase, بغض النظر عن *biġaḍḍ annaḍ̲ar ʿan* is translated into one equivalent word in English *‘regardless’*.

Can the core components of AMWE be substituted with other similar or synonym items? Several types of AMWE have a form of resistance to lexical substitutability or variations to their core essential parts, thus this criterion might be used as an indicator of potentially notable AMWE items. This can be seen in all the examples mentioned above of AMWEs.

More details about the conceptual framework and AMWE criteria are provided in sections 3.3 and 3.4.

4.7.1.2 Reliability testing and evaluation findings

Once the annotation guidelines were prepared and provided to the participants, corpus-based examples of each candidate were extracted that represented the actual use of potential AMWE in various linguistic contexts. This step was designed to enhance the manual annotation process and enable the coders to achieve the best possible outputs. Moreover, the annotators were free to consult corpus tools to discover the in-depth meaning of each expression and then classify them as true or false candidates. The annotators were also asked to select a list of good examples for each true candidate to add them later to the AMWEL. Table 4.19 presents several AMWE items along with their corpus-based instances.

Table 4.19: AMWEs and their corpus-based examples.

AMWEs		Corpus Example
يقشعر لها الجسم	yaqšaʿirr lahā aljism	- بودي ان اقول مهمتي ليست سهلة اتلقى يوميا من التهديدات و الشتائم ما يقشعر ل ها البدن و لكن ب المقابل الرسائل المشجعة التي تصلني
		- فاقض ما انت قاض انما تقضي هذه الحياة الدنيا " تلك المقولة التي تقشعر لها ابدان الطغاة و تطمئن بها قلوب الخائفين المتوجسين من الظلمة
		- و بينما ك نا نتحدث مع افراد العائلة سمعنا اصواتا يقشعر لها الجسم و حاصرتنا الحجارة من كل جهة
قرة عيني	qurratu ʿaynī	- و حبب الي النساء و الطيب , و جعل ف ي الصلاة قرة عيني
		- فقالت و اين انتي يا قرة عيني من هذه الكلمة
		- هم قرة عيني و انا اربي هم احسن تربية و اعمل و اشقى ل اوفر ل هم سبل العيش الكريم
كلام الليل يمحوه النهار	kalām allayl yamḥūh annahār	- و التحالفات تتغير بلمح البصر و ينطبق عليها القول الماثور " كلام الليل يمحوه النهار " و ل يذهب الشعب الى الجحيم
		- لكن تبين ان ك ل ذلك كان من قبيل كلام الليل يمحوه النهار , و بقيت دار ل قمان على حال ها
		- و يا هول ما رايت عند التصويت و اذا ب كلام الليل يمحوه النهار و اذا ب الشمس تكذب الغطاس
على جناح السرعة	ʿalā jināḥ assurʿa	- ثم هرب الملك على جناح السرعة الى طبرق ف ي حماية القاعدة البريطانية هناك
		- فتعالوا ب نا ل نرحل سويا على جناح السرعة ل نعب ش مع الحبيب المصطفى
		- الفنانة صباح دخلت اول من امس الى مستشفى قلب يسوع ف ي الحازمية و على جناح السرعة

Due to the time constraints, the evaluation was conducted on samples from the generated lists of 17,382 items. The samples consisted of 6000 items divided into 12

datasets and classified according to their head-words and the number of components of the expressions, as shown in Table 4.20.

Table 4.20: Basic information on the test datasets.

TS. Nu.	AMWE class	No. of components	TS. Nu.	AMWE class	Nu. of components
TS1	Nominal	2	TS7	Nominal	4
TS2	Prepositional	3	TS8	Prepositional	2
TS3	Verbal	2	TS9	Nominal	5
TS4	Nominal	3	TS10	Prepositional	4
TS5	Verbal	3	TS11	Prepositional	5
TS6	Verbal	4	TS12	Verbal	5

The candidates included in the evaluation task were selected randomly from the extraction outputs and reflected various structures and frequency levels of the extracted lists.

In the annotation task, the 12 datasets were evaluated by three teams of two judges who were trained linguists with experience in Arabic linguistics. To ensure the reliability of the annotation task and that an acceptable degree of inter-agreement was reached between the coders in the manual annotation, the annotation task was implemented several times on training samples consisting of 130 various AMWE candidates. In the training rating exercises, the participants were first introduced to the research project and then provided with detailed annotation guidelines. However, during training, an improvement was observed in the performance of manual classification and the annotators in the pilot evaluation task developed an understanding of the annotation tasks. Following the training annotation exercise, the team of annotators were asked to classify the test datasets into true or false AMWE; the first category represents notable AMWE candidates based on the annotation guidelines while the second category represents the candidates that cannot be considered valid AMWEs. Table 4.21 presents examples of the annotation exercise implemented in the evaluation.

Table 4.21: An example of the inter-rating annotation exercise[49].

-Tick all the phrases that you consider to be true AMWE candidates which can be stored in a lexicon.

- Use the provided examples or a corpus concordance tool or/and Arabic dictionaries if you need to understand the meaning of AMWE candidates in various context.

- If you are hesitant, you can make notes about this in the comment column.

No	AMWE	Type	Freq	True (1)	False (0)	Comments
1	من خلال	P	59900			
2	أكثر من	A	54114			
3	السلام عليه	N	39907			
4	من اجل أن	P	37889			
5	بالنسبة لـ	P	31243			

4.7.1.3 Measuring inter-coder agreement

The degree of inter-coder agreement was tested using the kappa statistic κ (Cohen, 1960). This is a test used to validate the null hypothesis H0 that the observed agreement is entirely due to chance. In other words, that the annotation is not reproducible. The kappa statistic is defined as the observed proportion of agreement minus the expected percentage of chance agreement $p_0 - p_c$ scaled to a standard range. The value of the test ranges from 0 to 1 where the higher the value, the better the agreement between the inter-coders, as can be seen in Table 4.22.

$$k = \frac{p_0 - p_c}{1 - p_c}$$

Table 4.22: Interpretation of the kappa agreement test's values (Viera and Garrett, 2005).

Kappa result	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

To measure inter-annotator agreement, the result of classifying a total of 1200 AMWE candidates was used which represented the 12 test datasets used in the evaluation task.

[49] More examples of test data and annotations test are provided in appendix G.

For each dataset, the raw agreement was calculated, followed by the κ score. Table 4.23 presents an example of a confusion matrix used to summarise the annotation findings for the first test dataset along with relevant statistical information.

Table 4.23: Summary of the two coders' manual annotation of TS1.

TS 1		Coder 1			
		False	True	Total	Percentage
Coder 2	False	40	11	51	51%
	True	4	45	49	49%
	Total	44	56	Total annotated items	100
	Percentage	44%	56%	No. of agreement	85
Pr(c)	50%	Pr(o)	85%	*k* test	0.70

Table 4.24 summarises inter-agreement statistics for all the test datasets used in the evaluation task in this experiment along with the overall averages for agreement information. This shows that the κ test values ranged from 0.4 to 0.8 in all the test datasets included in the annotation task.

Table 4.24: Agreement statistics for the 12 test datasets: raw agreement, κ score, and average.

Dataset	TS1	TS2	TS3	TS4	TS5	TS6	
Percentage of agreement	85	75	78	70	90	93	
κ test	0.7	0.5	0.6	0.4	0.8	0.8	
Dataset	TS7	TS8	TS9	TS10	TS11	TS12	Average
Percentage of agreement	74	72	91	79	84	78	75
κ test	0.5	04	0.8	0.6	0.7	0.6	0.6

These statistics suggest a moderate agreement which is an adequate measure of good reliability in the evaluation. Achieving a higher degree of inter-annotator agreement is difficult when classifying miscellaneous MWE items, even with the availability of detailed annotation guidelines and when conducting preliminary annotation exercises to enhance the overall reliability of the manual annotation tasks. Moreover, similar agreements have been found in corresponding studies in the literature, for instance Pecina (2009) reports an agreement of 0.49 using the Fleiss' κ test when measuring the inter-agreement of three coders, while Seretan (2011) found comparable κ test scores ranging from 0.49 to 0.60 based on various datasets.

Although several tasks can improve the manual classification of MWEs, this involves a substantial number of ambiguous items which might be interpreted from different perspectives by the annotators.

However, due to the absence of reference data, it was not possible to calculate the recall score in the evaluation for this experiment. Thus, precision scores were used to evaluate the performance of the AMWE extraction model based on the manual annotation of the 12 test datasets which represent the various morphosyntactic patterns applied in our extraction model. Precision in this instance was measured by the percentage of true extracted items divided by the number of all items included in the evaluation task, as shown in the following equation:

$$precision = \frac{true\ AMWEs}{all\ annotated\ AMWEs}$$

Table 4.25 shows a metric that summarises the annotation results of the 12 datasets. As mentioned previously, the total number of extracted items sampled in this experiment was 6500 candidates; however, after manual validation only 4557 validated AMWEs remained which constituted the AMWE lexicon and will also be used in subsequent research as reference data for evaluation purposes.

Table 4.25: Number of true AMWE items in the test sets based on manual annotation.

Dataset	TS1	TS2	TS3	TS4	TS5	TS6	
True positive items	442	430	427	415	409	396	
Dataset	TS7	TS8	TS9	TS10	TS11	TS12	Total
True positive items	386	377	307	343	330	295	4557

Table 4.26 shows the precision measures for the 12 data sets used in the evaluation experiments. These findings range from 0.57 to 0.77 with an average precision value of 0.72 for all the test datasets included in the evaluation.

Table 4.26: The precision values of the extracted 12 datasets.

Dataset	TS1	TS2	TS3	TS4	TS5	TS6	
Precision	0.88	0.86	0.85	0.83	0.82	0.79	
Dataset	TS7	TS8	TS9	TS10	TS11	TS12	MAP
Precision	0.77	0.75	0.61	0.69	0.66	0.59	0.70

These evaluation findings are in line with previous research that has implemented the MWE extraction model. For instance, Bounhas and Slimani (2009) achieved an overall precision value of 0.65 when extracting various types of AMWEs. Other MWE research (Seretan, 2011; Pecina, 2009; Moirón, 2005) have reported similar precision values ranging from 0.52 to 0.71 when extracting various types of MWE items.

4.8 Qualitative analysis

Thus far, the focus has been on the linguistic analysis and refinement tasks for the validated AMWE items from the evaluation test datasets. Overall, the findings of this experiment show that the AMWE extraction model works best with bigram and trigram candidates with precision scores above 0.8. This can be seen in Figure 4.9 which presents the dataset types and the extraction precision scores obtained by manual annotation of the test datasets.

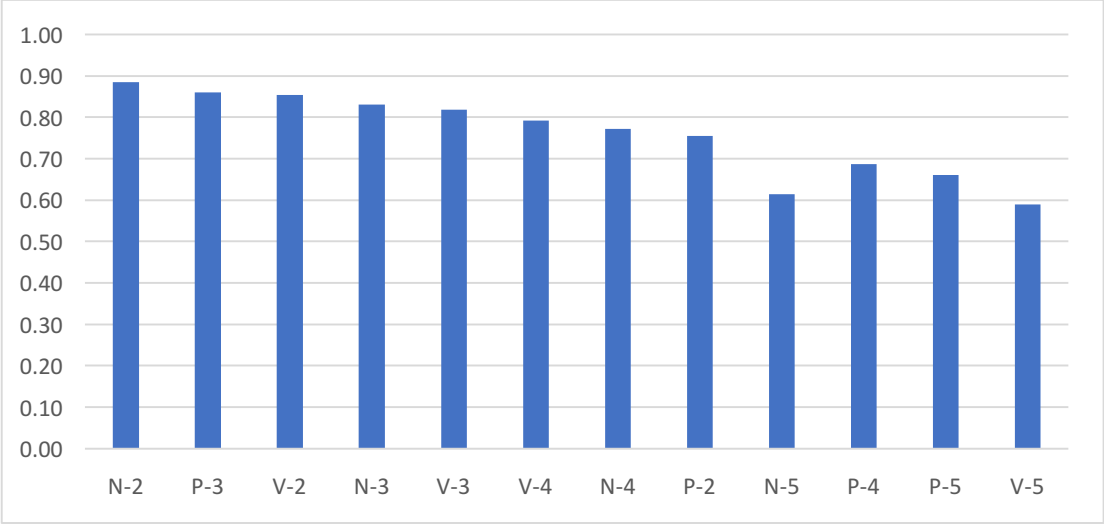


Figure 4.9: The extraction precision values for the test datasets.

Figure 4.10 presents the average precision scores for test datasets according to the lengths of the expressions. These show that the five-gram candidates have the lowest

precision score of 0.62 while the tri-gram candidates achieve the highest score of 0.84. These results are in line with previous MWE extractions where it was found that the retrieval of high-frequency sequences usually yields a better extraction output as the statistical methods work best with high-frequency rather than low-frequency items (e.g., Evert, 2005; Pecina, 2009).

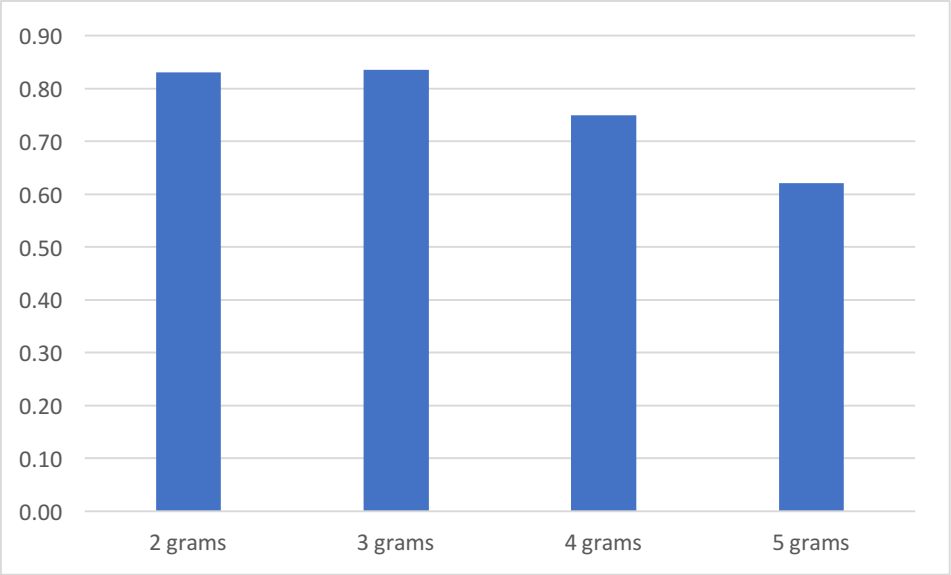


Figure 4.10: The average precision scores of the datasets based on the length of candidates.

Furthermore, linguistic analysis of the extracted items shows that they represent various types of syntactic constructions found in the literature (e.g., Najar et al., 2016; Al-Sabbagh et al., 2014; Meghawry et al., 2015; Hawwari et al., 2014). Table 4.27 presents various morphosyntactic structures of AMWE candidates with examples from the evaluation datasets.

As described in section 3.3.2, which explained the adopted definition of AMWE, there are no specific morphosyntactic constructs that are considered significantly more predictive in generating valid AMWEs. Instead, multiple patterns retrieve instances that emphasise the varying properties of AMWE at various levels of linguistic analysis. However, there is a need for more corpus-based evidence to support this statement through large-scale studies that cover an exhaustive listing of AMWE selection patterns.

Table 4.27: AMWE examples of various morphosyntactic patterns.

AMWE class	AMWE pattern	Instances	
Nominal	N-A	الآونة الاخيرة	alʾāwina alʾaḫīra
	N-P-N	النأي بالنفس	annaʾyu binnafs
	N-N-C-N	اهل الحل و العقد	ʾahl alḥalli wa alʿaqd
	N-N-C-N-N	احقاق الحق و ابطال الباطل	ʾiḥqāqu alḥaqqi wa ʾibṭālu albāṭil
Verbal	V-N	رفع المعنويات	rafʿ almaʿnawiyyāt
	V-N-N	تتواكب مع المتغيرات	tatawākabu maʿa almutaġayyirāt
	V-N-P-N	اختلط الحابل بـالنابل	ʾiḫtalaṭa alḥābilu binnābil
	V-P-Adv-N-N	يسعى من وراء هذا العمل	yasʿā min warāʾi hāḏā alʿamal
Prepositional	P-N	بالتالي	bittālī
	P-N-P	بالإضافة إلى	bilʾiḍāfa ʾilā
	P-N-N-P	بغض النظر عن	biġaḍḍi annaḏari ʿan
	P-N-N-A-A	بسم الله الرحمن الرحيم	bism allāhi arraḥmāni arraḥīm

The final linguistic analysis in this experiment related to the level of compositionality among AMWEs. The extracted AMWEs varied in their degree of idiomaticity; this means that the meaning of the AMWEs differed in relation to parts of the phrase; some phrases can be easily understood directly from their component parts, while others have a meaning that is irrelevant to component parts and are thus described as non-compositional or opaque MWEs. Mel'ćuk (1998) presented semantic classifications of phrases in terms of their degree of idiomaticity. The first category is full phrasemes, which is when the meaning of the phrase cannot be derived from its constituents. The second category is semi-phrasemes, which is when the meaning of the phrase matches the meaning of its components but has an additional meaning that is not related to its component parts. The third category is quasi-phrasemes, which is when the meaning of the expression derives directly from one part of the phrase and is partially or indirectly derived from another. Such degrees of semantic opacity degrees were found in the validated AMWE in this study. Multiple AMWE candidates were therefore classified into these three semantic categories, as shown in Table 4.28.

Table 4.28: Semantic opacity of the AMWEs.

	Semantic degree	Example
1	Full phrasemes	بالطبع, *biṭṭab* ʿ 'of course.'
2	Semi-phrasemes	إلى حد ما, *ʾilā ḥaddin mā* 'to somewhat.'
3	Quasi-phrasemes	السياسية الخارجية *assiyāsa alḫārijiyya* 'Foreign policy.'

The evaluation findings also show that the AMWEs represents various semantic domains, in line with the aim of this study which was to construct general reference lists of AMWEs. The semantic representations developed by Rayson et al. (2004) could be applied in the classifications of extracted items to enhance their usability in semantic-aware NLP applications. Several NLP applications benefit from the availability of semantically annotated LRs, such as the semantic tagger in the work of Rayson et al. (2004). Another application related to measuring the compositionality levels of MWE algorithms can be seen in the work of Piao et al. (2006) who utilised the semantic MWE lexicon of English to develop an automatic ranking model of MWEs based on their level of semantic idiomaticity

4.8.1 Error analysis

This section presents an analysis of the AMWE items that were classified as false candidates in the manual annotation task to explore potential sources of error in the retrieved lists. The following examples explain the nature of erroneous candidates. However, several types of error have been automatically removed from the initial extracted lists but, as expected, the outputs of automated processing were accompanied by multiple errors, as is often the case in ANLP processing tasks. The following list presents examples of sources of error:

The phrase involved an abbreviation, a proper noun, or numbers

Dialectical type of Arabic or foreign expressions.

Items appeared on the listing more than once because of different spellings or lexical variants of AMWE.

The phrases were meaningless and included clusters that consisted merely of articles or prepositions;

Named entity constructions such as proper noun or organisation names

Redundancy:items appeared repetitively with little variation due to errors in the linguistic analysis conducted in the extraction process.

Any AMWE candidates that did not meet at least one of the inclusion criteria based on our adopted conceptual framework.

Another source of errors was related to the result of automatic POS tagging and linguistic analysis implemented in the experiment. For instance, it was observed that, in many candidates, multiple clitics that should be separated were instead treated by the automatic tagger as one token, as shown in the following example:

حَرَاكٌ احْتِجاجِيٌّ عَلَى طُولِ البِلَادِ وَعَرْضِها *ḥarāk ʾiḥtijājī ʿalā ṭūl albilād waʿarḍihā*

where the two tokens (*ʾiḥtijājī* احتجاجي) and (*waʿarḍihā* وعرضها) should be tokenised as five tokens as follows:

احتجاج| ي | و | عرض | ها

ʾiḥtijāji | y | wa | ʿarḍi | hā

Furthermore, vice versa tokenisation errors can be observed in the following example:

مَلء الشُّقُوقِ فِ ي داخِل الخَلِيَّة

malʾ aššuqūqi fi y dāḫil alḫaliyya

In this example, the preposition في *fī* is analysed as two split tokens ي| ف which is a clear tokenisation error that might be attributable to the ubiquitous presence of the separated conjunction ف in the corpus. Other errors found in the task of POS tagging were those where the tool assigns the incorrect tag to the tokens. This can often be seen in more ambiguous POS classes such as adjectives and adverbs which have no clear clues in the text and therefore make it hard for the tagger to distinguish these POS classes in various linguistic contexts. Table 4.29 presents examples of several erroneous candidates and their associated errors.

Table 4.29: Examples of excluded AMWEs and the reasons for their exclusion.

AMWE candidates		Type of errors
مش عارف	miš ʿārifʿ	dialectical language
قطاع غزة	qiṭāʿ ġazza	NEs
ى الله علي ه و	aa ʾallāh ʿalī ha wa	meaningless construct
على سبيل المثال	ʿalā sabīl almiṯāl	Redundancy
بطولة ك اس العالم	buṭūla k as alʿālam	tokenisation error
على اعتبار ان	ʿalā iʿtibār anna	POS annotation error
الى المدينة	ʾilā almadīna	non-AMWE

4.9 Summary and conclusions

The result of this experiment yielded a refined list of 4557 AMWEs that met at least one of the manual evaluation criteria. The hybrid model adopted in this experiment utilised statistical and linguistic extraction methods for AMWEs that resulted in multiple lists of AMWE based on various morphosyntactic patterns. The manual evaluation of test datasets from the extracted list generated validated reference lists that can be used for the proposed lexicon of AMWEs and in different AMWE-aware ANLP tasks.

The extraction process began from the pre-processing stage which included the implementation of normalisation, automatic linguistic annotation, and the selection of morphosyntactic patterns. This was followed by statistical processing in which the n-gram model was used to generate multiple frequency-based lists of AMWE candidates. Sentence examples were provided for each list item to enhance usability and accessibility for the end-users of this resource. The selection of these examples underwent a qualitative analysis of randomly selected concordance samples from the corpus to determine the most frequent and relevant examples of selected AMWEs that represent various semantic senses of the expressions in multiple contexts.

Based on the evaluation findings presented in Figure 4.10, the most useful AMWE candidates were found in lengths that ranged from two to four components and will be the focus of the subsequent extraction experiments. The morphosyntactic patterns used in the study have generated predictive lists of AMWE candidates that also reflect

the variations of this phenomenon in SA at multiple levels of linguistic analysis. The semantic analysis of the AMWEs will enhance the utility of this list in different practical NLP tasks. For instance, knowing the non-compositional AMWEs enables an MT system developer to treat them as a single word, which ultimately increases the overall accuracy of the output of NLP applications.

This chapter has briefly reported on the implementation of a hybrid AMWE extraction model to extract various types of AMWE items to use in the development of an AMWE lexicon and as reference datasets for the following extraction experiments. In the model, the researcher drew upon the available state of the art Arabic linguistics disambiguation toolkits to implement various ANLP tasks and improve the overall findings of the experiment. The multiple processing phases applied to the corpus in this study have resulted in the discovery of multiple AMWEs that may be of great benefit for NLP and other language related tasks.

This experiment is the first step in a larger research project that aims to construct a comprehensive repository of AMWEs to assist in the process of integrating AMWE knowledge in relevant NLP applications. The subsequent AMWE extraction experiments reported in chapter five and six extend the current AMWE datasets by including less frequent AMWEs and special attention will be paid to the most predictive morphosyntactic patterns used in this experiment. The evaluation of various AMs in AMWE extraction will then be conducted to enhance the output of our extraction model; thus, the reference data extracted in this experiment will be improved and updated with new AMWE items in future research. Furthermore, in chapter 7 a detailed representation model will be described that is based on a comprehensive annotation AMWE scheme designed to represent various linguistic features of AMWE at multiple levels of analysis, including phonological, orthographical, syntactic, semantic and pragmatic features.

The extraction model in this experiment was also used to extract a small list of AMWEs for LP consisting of more than six hundred items. The pedagogical expressions listed have been used in English and other languages for a long time and have been found to be a beneficial pedagogical tool in multiple educational applications used for language learning.

5 Evaluation of Association Measures in AMWE Extraction

5.1 Introduction

In this chapter, a comparative evaluation will be presented of several AMs used in extracting bigram AMWEs. These are based on reference datasets developed in the previous study and described in chapter 4. This chapter reports the findings of four empirical experiments that used several AMs in the process of extracting and ranking lists of retrieved AMWE candidates. Before describing the results of the experiments in sections 5.4 to 5.7, a brief explanation of the AMs used in the evaluation tasks will be presented in section 5.2 and the methodology will be described in section 5.3. However, part of the work presented in this chapter has been published in Alghamdi and Atwell (2016a) and Alghamdi and Atwell (2016c).

5.2 Statistical association measures

AMs are multiple types of mathematical formula that calculate the association score between two objects in a corpus based on the frequency information. The initial use of AMs has been found in the information retrieval research field, but several later studies on AMWE found these statistical measures can also be used as practical statistical methods for extracting various lexical sequences from the text. More details about AMs are presented in section 2.2.2. However, in his comprehensive study on AMs, Evert (2004, pp. 76–77) classified AMs into four main categories based on their approach towards statistically measuring the level of associations between two events. These are summarised along with AM examples in Table 5.1.

Table 5.1: Major approaches to measuring associations (Evert, 2004, pp. 76–77).

Approach	Explanation	AM examples
The significance of association is derived from statistical hypothesis tests	Quantifies the amount of evidence that the observed sample provides against the non-association of a given pair type	binomial test, Poisson test, Fisher's exact test multinomial-likelihood, binomial-likelihood, Poisson-likelihood, the Poisson-Stirling
The degree of association	Estimates one of the coefficients of association strength from the observed data.	MI (mutual information MS (minimum sensitivity) Dice coefficient
Measures from information theory	Based on the information-theoretic concepts of entropy, cross-entropy, and mutual information.	MI-conf
Heuristic measures	Combines sample values that are considered to be good indicators of (positive) associations in various ways.	t-score MI2 MI3

In the experiment, the performances of several AMs were measured and are presented with their formulas in Table 5.2. The selection of these AMs was based on the analysis of several MWE extraction research studies that evaluated the performance of these methods and found encouraging results for those AMs used in the current study. Because the datasets and the language have a substantial impact on the performance of these AMs, it was useful to conduct several evaluation experiments to examine the performance of these AMs in a different experimental setting. For instance, Evert (2008, p. 31) states that:

'while some measures have been established as de-facto standards, e.g. log-likelihood in computational linguistics, t-score and MI in computational lexicography, there is no ideal association measure for all purposes. Different measures highlight different aspects of collectivity and will hence be more or less appropriate for different tasks.

Thus, in the experiments reported in this chapter, the performance of these AMs on several datasets were evaluated to establish which were the best to use to enhance the AMWE extraction model in the retrieval of multiple AMWE bigrams.

Table 5.2: Algorithms used to measure the association strength of the word pairs.

AMs	References	Formula
T-score	(Church et al., 1991)	$\frac{f_{xy-\frac{f_y f_b}{N}}}{\sqrt{\int \mathrm{xy}}}$
mutual information (MI)	(Church et al., 1990)	$log_2 \frac{f_{xy} N}{f_x f_y}$
MI3	(Daille, 1994)	$log_2 \frac{\int_{xy}^{3} N}{f_x f_y}$
MI.log_f	(Rychlý, 2008)	$MI - score \times log_{xy}$
logDice	(Rychlý, 2008)	$logDice = 14 + log_2 D = 14 + log_2 \frac{2fxy}{fx + fy}$
Log-likelihood(L.LK)	(Dunning, 1993)	$-2 \sum \mathrm{ij} \int ij \log \frac{f_{ij}}{f_{ij}}$
Minimum sensitivity (MS)	(Bruce and Pedersen, 1996)	$MS = min\left(\frac{O_{11}}{R_1}, \frac{O_{11}}{C_1}\right)$

5.3 Evaluation methodology

As mentioned in sections 2.2.4 and 4.7, there is no consensus in the literature regarding the optimal method for evaluating MWE extraction tasks. Nevertheless, most research takes advantage of evaluation techniques found in related research areas such as information retrieval and attempts to implement these methods on various related NLP evaluation tasks. In the current evaluation of different AMs, MWE extraction was considered a classification task where the best AM was the best predictor of correct MWE items in the datasets. The findings are illustrated in Table 5.3 in the form of a matching matrix which contains all the information related to the classification result as can be seen.

Table 5.3: Matching matrix showing the findings of the MWE classification task.

Actual items		**Predicted Items**	
		MWE	Non-MWE
Reference datasets	MWE	TP	FN
	Non-MWE	FP	TN

Based on the classification[50] findings in the contingency tables, the recall and precisions scores can be calculated based on the following formulas:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Precision = \frac{\text{T P}}{\text{T P} + \text{F P}}$$

$$Recall = \frac{\text{T P}}{\text{TP} + \text{FN}}$$

Regarding the reference data, in the evaluation experiments three bigram datasets developed in a study based on the AMWE extraction model were reported in chapter four of this thesis. The extracted items by AMs will be evaluated against these three reference datasets in the first three investigations. Table 5.4 presents AMWE random examples from the three datasets. Regarding the n-best list, which represents the highest ranked AMWE candidates when a specific AM is applied to the database in the literature, no optimal number of items included in the extracted list has been suggested. Thus, in the current research, the size of the extracted list was 100, which is in line with several previous studies in this area.

[50], The four classes can be described as follows:

- True Positives (TP): candidate was positive and predicted positive.
- False Negatives (FN) candidate was positive but predicted negative.
- False Positives (FP) candidate was negative and predicted positive.
- True Negatives (TN) candidate was negative and predicted negative.

In the evaluation experiments the same procedures were followed and can be summarised as follows:

Prepare the dataset used in the evaluation task.

Determine a frequency threshold of 10 per million words.

Select the search window span of -5 and +5 words.

Apply the AMs and retrieve several AMWE candidates list based on each AM.

Rank the extracted items in different lists based on their AM scores.

Compare the extracted list with the reference lists and classify the retrieved items into true or false candidates.

Calculate the AP and also the mean AP (MAP) values for each AM and present the evaluation findings.

In AM evaluation, precision-recall curves are usually used to estimate the performance of each AM based on the classification result of a random data sample. The overall interpretation of this curve is that the higher it is, the better the expected performance of a particular AM in extracting valid AMWE candidates. The use of these measures is not without its limitations and these are mainly related to the problematic measures of statistical difference between the precision-recall curves of various AMs. Given the drawbacks of only using precision-recall curves as an evaluation measure, especially in the absence of a large-scale reference dataset, the MAP scores in the experiments were used as the central evaluation figure to estimates the overall performance levels of AMs applied to each evaluation dataset. Furthermore, the significance was test used in the experiments, particularly the paired Student's t-test, to examine the statistical significance of the difference between the performance of AMs on various evaluation datasets.[51]

[51] In the evaluation experiments reported in this chapter, use was made of several AM tools such as (Kilgarriff et al., 2014), the UCS toolkit (Evert, 2004), and the Lancaster University corpus toolbox (Brezina et al., 2015).

Table 5.4: AMWE examples from three datasets.

Dataset 1		Dataset 2		Dataset 3	
muḥāwala yāʾisa	مُحَاوَلَةُ يَائِسَةُ	jarat alʿāda	جَرَتِ الْعَادَةُ	bisabab	بِسَبَب
masīra ḥāšida	مَسِيرَةٌ حَاشِدَةَ	ṣaʿid alminbar	صَعِدَ الْمِنْبَرَ	bism	بِاسْمِ
nisba ḍaʾīla	نِسَبَةٌ ضَئِيلَةٌ	ašārat aṣṣaḥīfa	اشارت الصَّحِيفَةُ	ʿalā ʾaʿtāb	عَلَى أعتاب
ʾāḏān ṣāġiya	آذانٌ صَاغِيَةٌ	šann alḥarb	شَنَّ الْحَرْبَ	ʾilā alʾabad	إِلَى الْأَبَدِ
qanāʿa tāmma	قَنَاعَةٌ تَامَّةٌ	ʾaḍāfat aṣṣaḥīfa	أضَافُت الصَّحِيفَةُ	ʿalā ʿātiq	عَلَى عَاتِقِ
ʾahdāf naḏīfa	أهدافٌ نَظِيفَةُ	ṣaḥḥ attaʿbīr	صَحَّ التَّعْبِيرُ	ʿalā ʿaks	عَلَى عَكْسِ
kalām fāriġ	كَلَامٌ فَارِغ	tajāwuz alḥudūd	تَجَاوُزُ الْحُدود	binnisba	بِالنِّسْبَةِ
šakl jaḏrī	شَكْلٌ جَذْرِي	tanāwul alluḥūm	تَنَاوُلُ اللُّحُوم	biliʾiḍāfa	بِالإضافَة
diqqa mutanāhiya	دقة متناهية	ʾaḍāfat almaṣādir	أضافت الْمَصَادَر	bittālī	بِالتَّالِي
makāna marmūqat	مَكَانَةٌ مَرْمُوقَةٌ	aqtaḍat aḍḍarūra	اقْتَضَتِ الضَّرُورَة	bilfiʿl	بِالْفِعْلِ

Table 5.4 shows AMWE items from the three-datasets used in the evaluation tasks. The first data set represents various types of nominal AMWE bigrams that cover different semantic domains, the second list includes set of verbal bigrams that provide for multiple types of support verb constructions, and dataset 3 shows various kinds of prepositional AMWE that are mainly used as discourse markers in different linguistic contexts. The main reason for measuring the performance of AMs on multiple datasets is because most MWE extraction research using AMs found that they are usually sensitive to the type of dataset used in the evaluation tasks. Thus, to determine the most predictive AM for several kinds of AMWEs the evaluation experiments have to performed on multiple datasets. Table 5.5 presents an example of the n-best list of extracted AMWE candidates ranked in descending order based on the MI scores.

Table 5.5: Examples of AMWE candidates ranked in descending order based on MI.

AMWE candidates		MI scores
hazza ʾarḍiyya	هزة أرضية[52]	13.73
diqqa mutanāhiy	دقة متناهية	12.56
bīʾa jāḏiba	بيئة جاذبة	12.51
furṣa sāniḥa	فرصة سانحة	12.38
biyʾat ḫaṣba	بيئة خصبة	11.90
ʾarḍiyya muštaraka	أرضية مشتركة	10.65
makāna qudsiyya	مكانة قدسية	9.55
diqqa rawʿa	دقة روعة	8.27
diqqa yuʿālij	دقة يعالج	8.16
biyʾat taḥsīn	بيئة تحسين	8.02
makānat qulūb	مكانة قلوب	7.85
biyʾa mujtamaʿ	بيئة مجتمع	7.35
makānat almarʾa	مكانة المرأة	7.09
furṣa kay	فرصة كي	6.11

5.4 Experiment 1

This experiment involved conducting a comparative evaluation of the use of several AMs in extracting nominal AMWE bigrams. The following subsections briefly report the experiment applied to dataset 1, section 5.4.1 highlights the main experimental setting, and section 5.4.2 describes with examples the dataset used in this evaluation task. Section 5.4.3 then illustrates the procedures followed and, finally, sections 5.7.5 and 5.7.6 describe and discuss the core findings and provide a summary of this evaluation experiment.

5.4.1 Experimental setting

In this experiment, the ArTenTen corpus described in section 4.3 was adopted in the development of the reference lists. The corpus was automatically POS annotated using the SAP toolkit and covered a wide range of SA varieties and semantic domains. The AMs were applied to dataset one by extracting lists of AMWE surface bigrams and ranking them in multiple tables based on the scores of each AMs. Based on the

[52] The true candidates are represented in an underlined font.

reference list, the AMWE candidates included in each n-best evaluation list were classified, following which, based on the average precisions, the MAP score for each AM was calculated which is the primary evaluation figure indicating the overall performance of AMs applied to the dataset. Based on random samples from the evaluation lists, precision-recall curves will be presented to show the best performing AMs in this experiment.

5.4.2 Dataset

The dataset used in this experiment consisted of several types of nominal bigram extracted from the SA corpus, as shown in Table 5.6 which presents examples of AMWE constructions from dataset 1.

Table 5.6: Nominal structures and their instances from dataset 1.

Structures	AMWE examples	
DTNNS_DTJJ[53]	alquwwāt almusallaḥa,	القوات المسلحة
	aljihāt almuḫtaṣṣa,	الجهات المختصة
	almašrūbāt alġāziyya	المشروبات الغازية
NN_JJ	šakl mubāšir,	شكل مباشر
	tāra uḫrā,	تارة اخرى
	ġayr masbūq,	غير مسبوق
NN_DTJJ	murūr alkirām,	مرور الكرام
	ḏāt albayn,	ذات البين
	dimāʾ al’abriyāʾ,	دماء الابرياء
NN_JJR	niṭāq awsaʿ,	نطاق اوسع
	ahammiyya quṣwa,	اهمية قصوى
	dawla ʿuḏmā,	دولة عظمى

[53] This POS notation was based on SAP toolkit tagset which can be found in Appendix B of this thesis.

NNP_JJ	ʿibāra ʾuẖrā ,	عبارة أخرى
	aṯar rajʿī,	اثر رجعي
	bawtaqa wāḥida,	بوتقة واحدة
DTNN_JJ	alʿaks ṣaḥīḥ,	العكس صحيح
	alfurṣa sāniḥa,	الفرصة سانحة
	alḥāja māssa,	الحاجة ماسة
NN_NNS	qāʿidat bayānāt,	قاعدة بيانات
	talbiyat iḥtiyājāt,	تلبية احتياجات
	ittiẖāḏ ijrāʾāt	اتخاذ اجراءات

Nominal expressions are one of the most dominant class of AMWE; these include multiple syntactic structures and lexical variants that can be seen in the examples of constructions found in the reference list which covers a wide range of AMWE nominal bigrams. Thus, it is therefore useful to measure the performance of AMs on this dataset.

5.4.3 Performing the experiment

Following the procedures of the experiment described in section 5.3, several lists of surface nominal and open class bigram candidates were extracted based on various statistical AMs. The top 100 candidates were then ranked for each AM used in this experiment which resulted in 7 ordered lists of potential AMWEs. Table 5.7 presents instances from the ranked candidate lists using the 7 AMs applied to dataset 1.

Table 5.7: Examples of AMWE candidates extracted by 7 AMs applied to dataset1.

AM	Candidate Examples		Score
T-score	رضى الله	raḍī ʾallāh	47.158
MI	مكانة مرموقة	makāna marmūqa	12.246
MI3	دقة فائقة	diqqa fāʾiqa	20.43
L.Lk	مصادر أمنية	maṣādir ʾamniyya	40.344
MS	بيئة نظيفة	bīʾa naḍīfa	0.01682
Log.Dice	أرضية مشتركة	ʾarḍiyya muštaraka	9.17
MI.log.F	مجال العمل	majāl alʿamal	79.822

The top 100 candidates for each AM applied to dataset 1 were then evaluated and the ranked lists assessed in comparison to the reference lists, the MAP scores, and the precision and recall curves presented for the performance of AMs on the nominal dataset.

5.4.4 Results and discussion

The findings of the experiment on dataset 1 show a good overall result regarding AM performance implemented in this study. Figure 5.2 presents the MAP score for AMs applied to the dataset. The best method evaluated by the MAP score was MI which achieved more than 0.9 scores, followed by MI.LF and L.LK which obtained MAP scores of more than 0.8. T-score and MI3 were ranked as the lowest performing AMs in this experiment with MAP scores above 0.5.

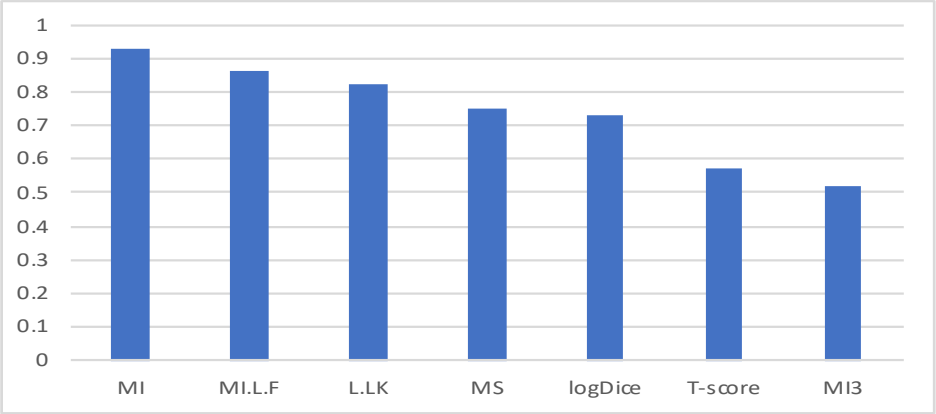


Figure 5.1: The MAP scores of AMs applied to dataset 1.

However, a baseline method based on a random ranking of AMWE candidates would achieve a MAP score of 0.23. Figure 5.2 presents precision-recall curves for the highest performing AMs applied in this experiment. The curves show an estimation of the three AMs' performance based on a random sample from the n-best lists used in this evaluation.

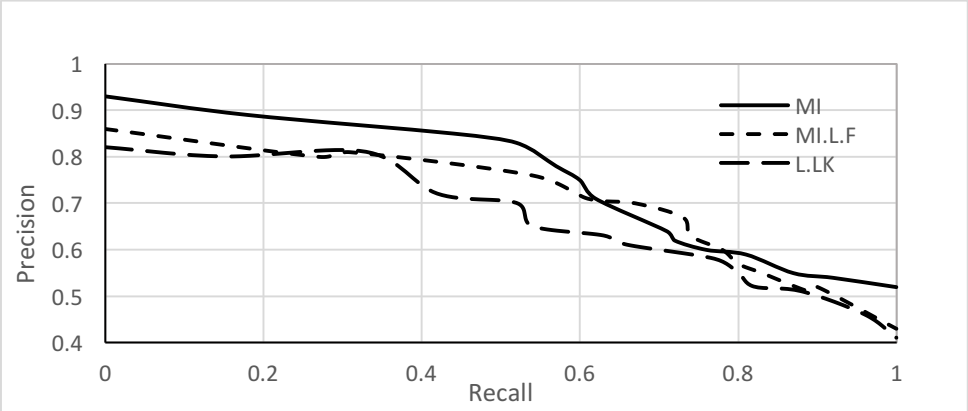


Figure 5.2: Precision-recall curves of the best 3 AMs applied to dataset 1.

The high performance of AMs on this dataset might be due to the dominant proportion of nominal AMWE which is by far the most common type of MWE in the language;

this aligns with the fact that AMs usually works better with a frequent string (Evert, 2004). Overall, this result was generally in line with previous MWE research findings. For instance, in his evaluation of several AMs, Pecina (2009) found that the T-score obtains the lowest MAP score of under 0.3. MI, MI.L.F and L.LK AMs were also found to be the best predictors in extracting multiple types of MWE (e.g., (Evert, 2004; Moirón, 2005; Attia et al., 2010). Table 5.8 presents several examples from the true AMWE extracted by the three high-performing AMs

Table 5.8: Examples of true AMWEs extracted by the best AMs for dataset1.

AM	Candidate Examples		AM Score
MI	شكل متتالي	šakl mutatālī	10.240
	سبحان الله	subḥān ʾallāh	7.523
	مجال الطاقة	majāl aṭṭāqa	7.219
L.Lk	وجهة النظر	wijhat annaḏar	15,378.713
	بناء القدرات	bināʾ alqudrāt	4,704.295
	اطلاق الرصاص	iṭlāq arraṣāṣ	1,049.962
MI.log.F	الرأي العام	arraʾy alʿāmm	75.668
	الفكر المعاصر	alfikr almuʿāṣir	57.054
	الموقع الالكتروني	almawqiʿ alilikturūnī	55.778

5.4.5 Summary

Thus far, this experiment has evaluated the use of 7 AMs on the previously constructed list of a nominal AMWE bigram. The statistical tests were applied and evaluated against the reference list by classifying the top-ranking lists of various AMs. The finding show a good overall performance with the top AMs in this experiment being MI, MI, L.F, and L.LK, which is generally in line with several MWE evaluation experiments conducted on various evaluation datasets. Hence, in the following experiments it will be important to see whether these AMs yield a different performance evaluation when they are implemented on verbal and prepositional AMWE bigrams. These findings will be presented in sections 5.5 and 5.6.

5.5 Experiment 2

Following the same procedures conducted in experiment 1, another experiment is presented which implements a comparative evaluation of AMs in extracting verbal AMWE bigrams (dataset 2). The experimental setting is described in section 5.5.1, and reference dataset two is described briefly with examples in section 5.5.2. Section 5.5.3 outlines the main experiment steps, and then, in sections 5.5.4 and 5.5.5, the evaluation findings are presented along with examples and the core findings are discussed.

5.5.1 Experimental setting

This experiment was also applied to the ArTenTen corpus and the same procedures were used as outlined in sections 5.3 and 5.4.1. The n-best lists method was used in the evaluation of AMs based on the MAP scores which estimate the overall performance of AMs in extracting verbal AMWE candidates. The 7 AMs were applied to dataset 2, and various candidate lists were generated based on different AMs. The items in each list were ranked in descending order according to their AM scores.

5.5.2 Dataset

The reference list used in this experiment includes various types of AMWE bigram which represent various types of support verb expressions in SA. Table 5.9 presents examples of the verbal patterns found in dataset 2.

Table 5.9: Verbal structures with their instances from dataset 2.

Structures	AMWE examples	
VBN_NN	talʿab dawr	تلعب دورا
	rāḥ ḍaḥiyya	راح ضحية
	yaḥill maḥall	يحل محل
VBD_DTNN	asdal assitār	اسدل الستار
	fataḥ albāb	فتح الباب
	rafaʿ aḍḍulm	رفع الظلم

VBP_NNS	tulabbī iḥtiyājāt	تلبي احتياجات
	yuqaddim ḫadamāt	يقدم خدمات
	tuwājih taḥadiyāt	تواجه تحديات
VB_IN	taḥtawī ʿalā	تحتوي على
	taʿtamid ʿalā	تعتمد على
	tatʿāmal maʿ	تتعامل مع
VBD_NNS	iqtaḥamat quwwāt	اقتحمت قوات
	nālat istiḥsān	نالت استحسان
	rafaʿt rāsī	رفعت راسي
VBD_DTNNS	aṯbatat addirāsāt	اثبتت الدراسات
	hataf almutaḍāhirūn	هتف المتظاهرون
	rafaʿ alʿuqūbāt	رفع العقوبات

5.5.3 Performing the experiment

First, the 7 AMs were applied to dataset 2 and multiple ranked lists of AMWE candidates were retrieved which were then sorted in descending order based on AM scores. The procedures followed here are the same as in the previous experiment described in section 5.4.3. Table 5.10 presents several examples of extracted AMWE candidates in dataset 2 with their AM values following the application of several statistical tests.

Table 5.10: Examples of AMWE candidates extracted by AMs applied to dataset 2.

AM	Candidate Examples		AM Score
T-score	فتح الباب	fataḥ albāb	33.328
MI	حفز النمو	ḥaffaz annumū	9.422
MI3	يقع فريسة	yaqaʿ farīsa	22.993
L.Lk	يدخل الجنة	yadḫul aljanna	6,391.838
MS	تضم عددا	taḍumm ʿadad	0.02028
Log.Dice	يشارك البطولة	yušārik albuṭūla	7.295
MI.log.F	يسهم اسهاما	yushim ishām	36.341

5.5.4 Results and discussion

Following the evaluation of several n-best lists of AMWE candidates extracted based on AMs, the overall MAP score for each method was calculated. Figure 5.3 shows the

MAP scores for all AMs applied in this evaluation task. This shows that MI and MI.log.F achieved the highest MAP score in ranking verbal AMWE bigrams with a value above 0.8, while T-score appeared to be the lowest performing AM in this experiment with a score under 0.5. The remaining AMs achieved similar MAP scores, ranging from 0.62 for MI3 to 0.69 for MS statistical measure. A baseline method based on a randomly selected list of AMWE candidates in this experiment achieved a MAP score of 0.21.

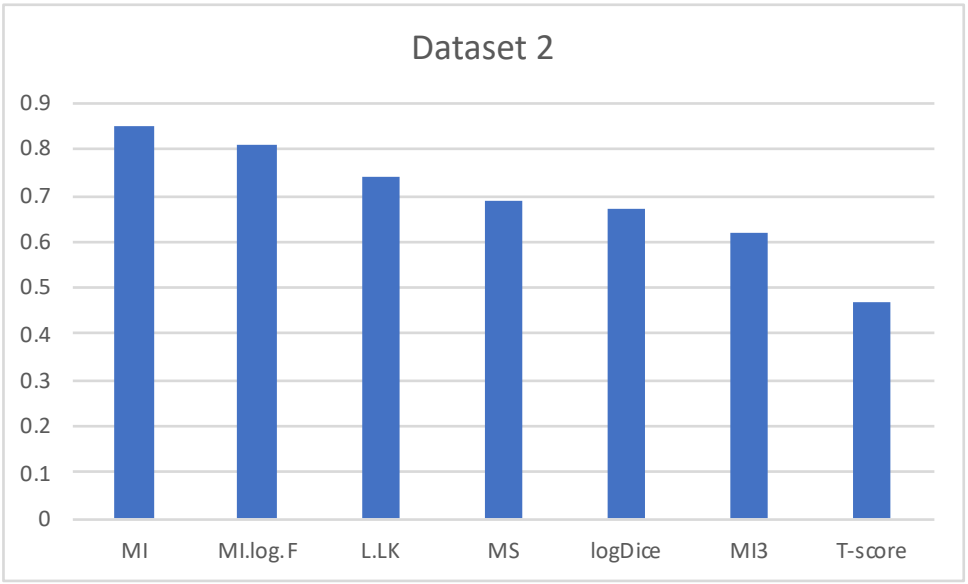


Figure 5.3: The MAP scores of AMs applied to dataset 2.

Figure 5.4 shows the precision-recall curves based on the recall and average precisions of the three best AMs when ranking AMWE candidates.

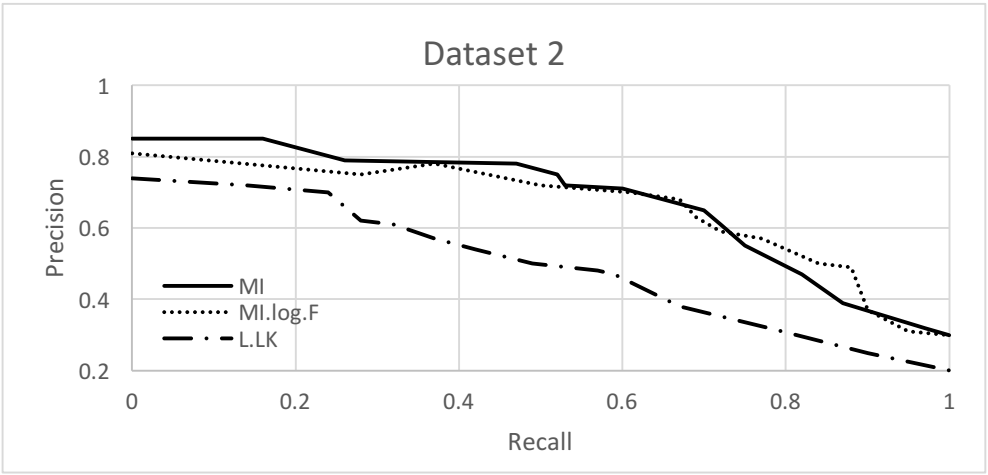


Figure 5.4: Precision-recall curves of the best 3 AMs applied to dataset 2.

As can be seen in the estimated performance curves, MI and MI.log.F exhibit a strong overlap in ranking AMWE candidates which indicates similar levels of precision and recall scores in comparison to L.LK which yields lower performance scores in classifying AMWE items. Table 5.11 presents several examples from the evaluated AMWE candidates ranked by the three best AMs in this experiment.

Table 5.11: Examples of true AMWEs extracted by the three best performing AMs on dataset 2.

AM	Candidate Examples		AM Score
MI	يبلغ أشده	yabluġ ʾašaddah	12.150
	يمثل تحديا	yumaṯṯil taḥaddiyan	10.282
	مجال الطاقة	majāl aṭṭāqa	9.659
L.Lk	يبلغ عدد	yabluġ ʿadad	8,098.407
	تتعامل مع	tataʿāmal maʿ	624.857
	يأتي بغتة	yaʾtī baġta	355.370
MI.log.F	تصب اللعنات	taṣubb allaʿnāt	28.607
	يتناسب طرديا	yatanāsab ṭardiyan	43.525
	يعيش منعزلا	yaʿīš munʿazilan	24.766

5.5.5 Summary

This experiment measured the performance of several AMs in extracting verbal bigrams based on the previously developed reference list. However, applying these AMs yields lower performance scores in comparison to their performance on dataset 1. One obvious explanation for this performance is related to the high frequency of nominal AMWE in the data which is one of the critical factors in improving the general performance of statistical measures. In the following experiment the same AMs will be applied on a different preoperational AMWE dataset (dataset 3) to explore the possible similarities and differences between AMs in ranking AMWE candidates on various datasets.

5.6 Experiment 3

In this section, the third comparative evaluation experiment will be reported which applies AMs to dataset 3, which mainly consists of multiple types of prepositional AMWE bigram. The primary objective of this experiment is to measure the performance of selected AMs in ranking AMWE candidates by following the same

procedures described in section 5.3. First, the nature of dataset three, which is used as a reference list in this evaluation task, will be illustrated with examples and then the experimental results will be reported by highlighting the best AM performance on this dataset. Several examples of true classified AMWEs in this study will also be presented.

5.6.1 Experimental setting

The setting applied to this experiment is similar to the previous two studies The ArTenTen corpus used in this experiment and the AMs used to retrieve a list of AMWE candidates were based on information frequency. The procedures described in section 5.3 were implemented based on the reference lists. From these. the performance of AMs in the n-best lists of AMWEs were evaluated.

5.6.2 Dataset

The reference dataset used in this evaluation experiment contains several kinds of AMWEs including main prepositions and other types of word class included in the reference dataset as described in detail in chapter 4. Table 5.12 presents some examples of AMWEs found in dataset 3 with their POS patterns which explain the variants of items in the reference list.

Table 5.12: Nominal structures with their instances from dataset 3.

Structures	AMWE examples	
IN-DTNN	binnisba	ب النسبة
	bi liḍāfa	ب الاضافة
	bi attālī	بالتالي
	bi arraġm	ب الرغم
IN-V	fīmā aʿlam	فيما اعلم
	fīmā yataʿllaq	فيما يتعلق
	fīmā yalī	فيما يلي
	fīmā yabdū	فيما يبدو

IN-R	fīma iḏā	فيما اذا
	bi hākāḏā	ب هكذا
	iḏ ṯammata	اذ ثمة
	iḏ ṯālamā	اذ طالما
IN-WRB	bi hayṯu	ب حيث
	ilā matā	الى متى
	ilā ayn	الي اين
	iḏ kayfa	اذ كيف
IN-NN	bi šakl	ب شكل
	ilā jānib	الى جانب
	ʿan ṭarīq	عن طريق
	bi sm	ب اسم
JJR-DTNN	aḍʿaf al'īmān	اضعف الايمان
	aḫar attaʿāzī	احر التعازي
	afḍal assubul	افضل السبل
	akṯar alaḥyān	اكثر الاحيان

5.6.3 Performing the experiment

The seven AMs were applied to extract several evaluation lists based on dataset 3 and the retrieved candidates were then sorted by their AM scores in descending order. Based on the AP scores, the MAP figure was then calculated for each AM where the main evaluation figure summarises the overall performance of each statistical test implemented in this experiment. Table 5.13 presents several examples from the extracted evaluation lists along with their AM scores.

Table 5.13: Examples of AMWE candidates extracted by AMs applied to dataset 3.

AM	Candidate Examples		Score
T-score	بالاضافة	bilʾiḍāfa	4.884
MI	من خلال	min ẖilāl	12.246
MI3	الأفضل ان	alʾafḍal an	22.486
L.Lk	من خلال	min ẖilāl	246.581
MS	الا اذا	illā iḏā	0.02853
Log.Dice	حيث تم	ẖayṯ tamma	8.410
MI.log.F	الجدير بـــ	aljadīr bi	38.735

5.6.4 Results and discussion

The finding of this experiment generally shows lower performance for all AMs in comparison to the outcome of experiments 1 and 2. However, this might be due to the nature of dataset three which represents prepositional and adverbial AMWEs which constitute a smaller proportion of AMWE. However, although most prepositional AMWEs have a high-frequency level, they are ultimately limited in number in languages which directly affects the number of MWEs related to them.

Figure 5.5 presents the MAP scores of AMs in descending order. These show that three AMs achieved a similar result of around 0.7 while MI3 and T-score were the least useful AMs in the ranking of AMWE candidates in this experiment. This finding is in line with previous MWE research which found that MI and L.LK are usually among the top AMs when extracting multiple types of AMs (e.g., Attia and Tounsi, 2000; Bounhas and Slimani, 2009)

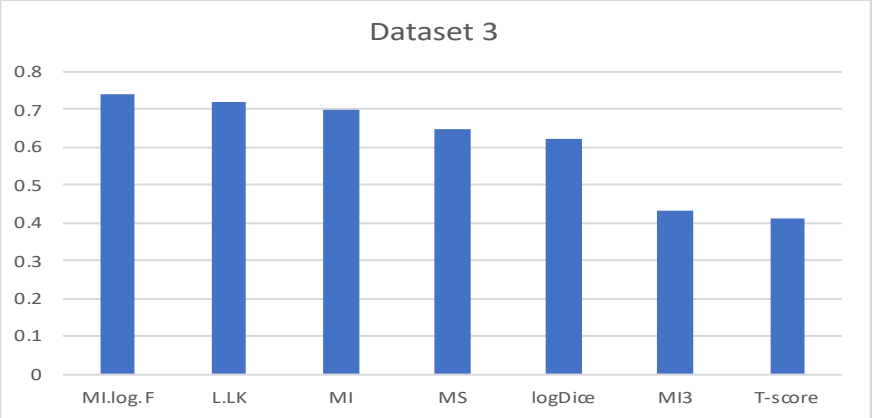


Figure 5.5: The MAP scores of AMs applied to dataset 3.

Figure 5.6 presents the precision-recall curves that reflect the performance of the three best AMs in this experiment. The MI.log.F AM slightly outperformed the MI and L.lk AMs, particularly within the low recall scores; however, there is an overlap between the performances of these three AMs in terms of high recall scores when applying these AMs to dataset 3.

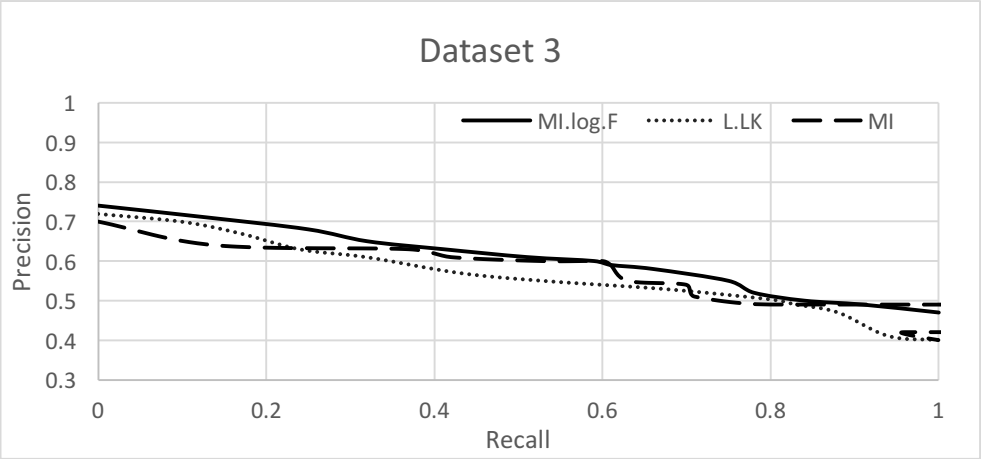


Figure 5.6: Precision-recall curves of the best 3 AMs applied to dataset 3.

Table 5.14 presents several examples of true AMWEs extracted by the three best performing AMs on dataset 3 in this experiment.

Table 5.14: Examples of true AMWEs extracted by the best performing AMs on dataset 3.

AM	Candidate Examples		AM Score
MI	كيف تحكمون	kayf taḥkumūn	10.168
	أيضا اقرأ	ʾayḍan iqraʾ	6.300
	ناهيك عن	nāhīk ʿan	4.338
L.Lk	وهكذا	wahākāḏā	60,087.377
	لطالما	laṭālamā	4,409.179
	هنا تكمن	hunā takmun	3,926.880
MI.log.F	هكذا دواليك	hākāḏā dawālīk	65.710
	على الاقل	ʿalā al’aqall	59.507
	طالما ردد	ṭālamā raddad	24.533

5.6.5 Summary

This experiment evaluated the performance of seven AMs in ranking AMWE candidates on dataset three which consisted of multiple types of prepositional and other types of AMWE. However, the overall results show lower performance for all AMs in comparison to their corresponding performances on datasets 1 and 2. The result of these three experiments provide evidence that the three AMs of MI, L.Lk and MI.log.F achieve the best performance in classifying AMWE candidates. Thus, in the AMWE extraction model, these AMs should be used to enhance statistical processing by generating multiple types of AMWEs based on AM tests. However, several factors

involved in the experimental setting might affect the overall results, such as the size of the reference data, the number of extracted n-best items, and the selected AMs. These should be borne in mind when interpreting the findings of this study and when attempting to apply these AMs in different AMWE extraction contexts.

5.7 Experiment 4

This experiment implemented what is referred to as the collocation of the Arabic keywords approach to extracting AMWEs in the form of high frequency but semantically regular formulas that are not restricted to any syntactic construction or semantic domain. The study applied several distributional semantic models to automatically extract relevant MWEs related to Arabic keywords. The datasets used in this experiment were rendered from a newly developed corpus-based Arabic wordlist consisting of 5,189 lexical items that represent a variety of SA genres and regions. The new wordlist was based on an overlapping frequency arising from a comprehensive comparison of four large Arabic corpora with a total size of over 8 billion running words. Empirical n-best precision evaluation methods were used to determine the best AMs for extracting high frequency and meaningful MWEs. The gold standard reference MWE list was developed in previous studies and manually evaluated against well-established quantitative and qualitative criteria. The results demonstrate that the MI.log_F AM achieved the highest results in extracting significant AMWEs from the large SA corpus, while the T-score association measure achieved the lowest results.

5.7.1 Introduction

Extracting the most common and meaningful MWEs associated with a frequency based Arabic wordlist - the primary concern in this study - is the basis for a useful LR that can be used in various language-related applications. The current study uses high frequency and significant AM scores as reliable predictors of a list of useful MWEs

Because the linguistic units extracted in this study were not restricted to any syntactic construction or semantic domain, the term MWEs was used as an umbrella to refer to various types of linguistic units in general. Thus, the current study adopts a practical definition of Arabic MWEs which primarily concentrates on any syntactic

construction from different language domains that make high frequency use of semantically regular phrases.

This is a preliminary study to explore a range of well-known AMs in extracting meaningful and high-frequency Arabic MWEs from a large SA corpus. The primary objective of this evaluation experiment is to determine the most reliable AM which can then be used as a predictor for the right collocates of the lexical items derived from a corpus-based Arabic wordlist.

5.7.2 Experimental setting

Association scores were used to rank the MWEs candidates extracted from a large corpus and precision scores were computed for the sets of n-highest-ranking. Thus, the first step in this experiment was to prepare a gold standard list of MWEs. For this, an AMWE list from a previous study was adopted, as described in chapter 4 of this thesis. In this experiment, six types of well-known AMs were selected: t-score, mutual information (MI), MI3, logDice, MI.log_f and L.LK. Table 5.2 presents the equations for these AMs along with their references.

5.7.3 Datasets

Two datasets comprising 50 high and low-frequency lexical items were selected for this experiment. The words in these datasets were extracted from a newly developed corpus-based wordlist of the most frequent SA words, based on their overlapping frequency and dispersion in a comprehensive comparison of four large SA corpora of over 8 billion running words, with the final wordlist consisting of more than 5 thousand items. The new list was automatically lemmatised and morphologically analysed using the MA toolkit illustrated in section 4.4.2. Figure 5.7 shows the distributions of word classes in the new Arabic wordlist.

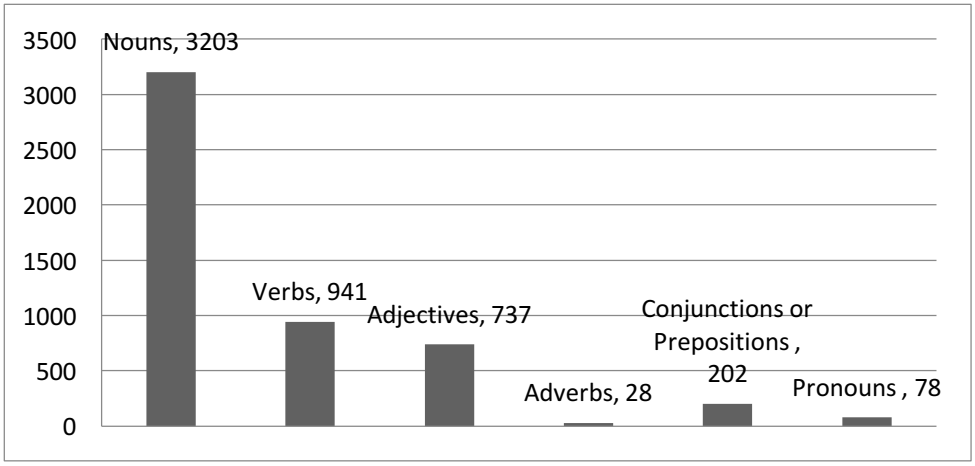


Figure 5.7: Distribution of word classes in the new corpus-based Arabic wordlist.

Each word in the dataset has an equivalent MWE from a previously developed gold standard MWE list. The reason for dividing the data set into high and low-frequency samples is to measure the node word frequency effect on the performance of AMs. Tables 5.15 and 5.16 show the five highest and lowest node words used in this experiment, along with their overlapping frequencies.

Table 5.15: The five highest node words.

Words	POS
من min 'from'	prep
على ʿalā 'on'	prep
هٰذا hāḏā 'this'	pron
خاصة ẖāṣṣa 'private'	verb
يوم yawm 'day.'	noun

Table 5.16: The five lowest node words.

Words	POS
التنافس attanāfus 'competition'	noun
قاسية qāsiya 'severe'	noun
مدرج madraj 'runway'	noun
يستلزم yastalzim 'require'	verb
حصانة ẖaṣāna 'immunity'	noun

5.7.4 Performing the experiment

The study was conducted in two rounds comprising the high and low-frequency data sets, each using the same procedures in the following steps. First, a threshold with a

minimum frequency of 10 per million was selected within a search window of two to four words, and the six AMs were then computed for each node word. The highest identified collocates were recorded and ranked based on different AMs, with the precision of each node word calculated as shown in the following equation:

$$precision = \frac{attested\ FSs}{all\ extracted\ FSs}$$

The average precision (AP) for each AM was then calculated for each node word and, finally, the mean average precision (MAP) for each AM was calculated for all node words. The experiment was performed on the ArTenTen SA corpus which consists of more than 7.4 billion running words.

5.7.5 Results and discussion

Figure 36 shows the MAP scores for each AM using the high-frequency data set in the first round of this experiment. This shows that the MI.log_f and MI measures achieved the highest MAP scores with a MAP score of over 0.85, while the t-score and MI3 were the least useful scores in terms of identifying MWEs among the high-frequency lexical items, with MAP scores below 0.50. The logDice and the L.LK achieved good scores in predicting the correct sequences with a MAP score of over 0.50. Overall, most AMs used with this data set achieved moderate to high MAP scores, except the T-score with a score of below 0.50. This result aligns with that of Alrabiah et al. (2014) who found that the MI.log_f score outperformed other AMs in predicting the lexical collocations in small and large CA corpora. However, other studies on Arabic collocations have found that the L.LK was the best AM in extracting lexical collocations (e.g.,Boulaknadel et al., 2008; Saif and Aziz, 2011), although these studies did not use the MI.log_f in their evaluation of AMs. This factor, along with the different experimental setting, might explain the variations that arose when determining the best AMs in the current experiment.

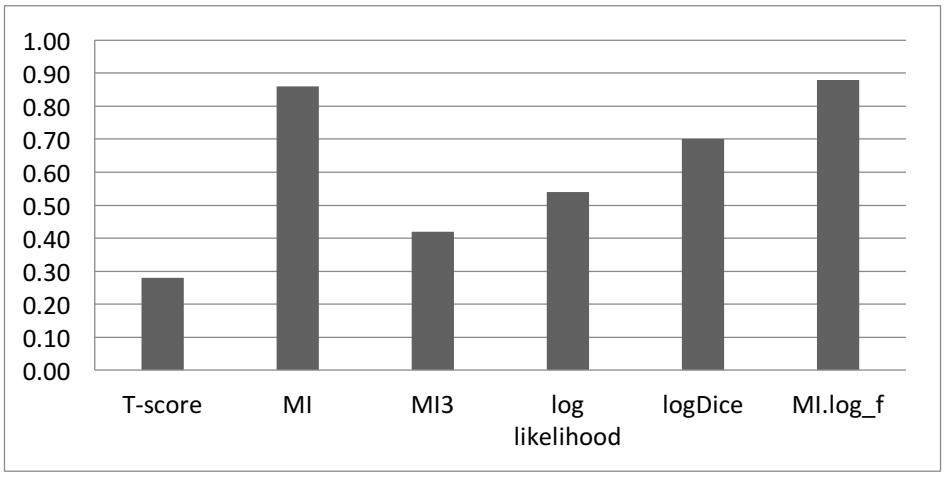


Figure 5.8: MAP scores of the AMs for the first dataset.

In the second round of the experiment, where the least frequent lexical items were used as the node words in MWEs extraction, the MAP scores in Figure 5.9 show an overall drop in the performance of most AMs. This is because most AMs usually work better with high-frequency data. In addition, the MI.log_F and the logDice outperformed other AMs with a MAP score of over 0.75. This suggests they are the best AM predictors when it comes to extracting the collocation of less frequent node words.

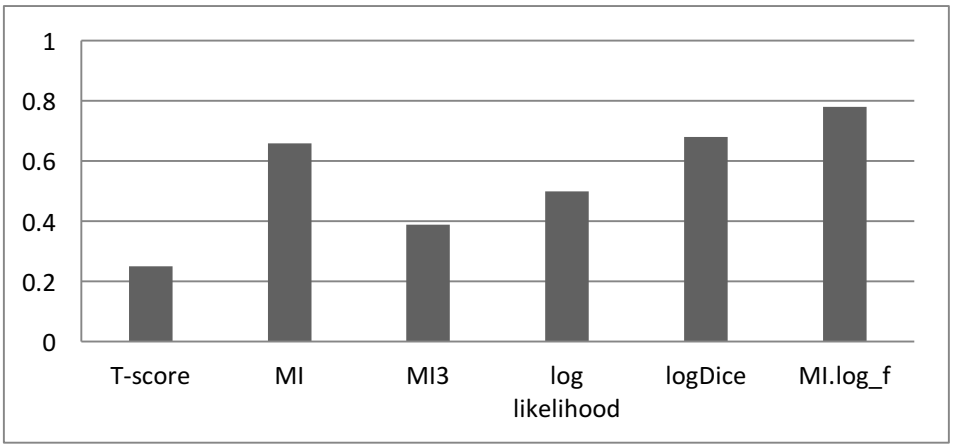


Figure 5.9: MAP scores of AMs for the second dataset.

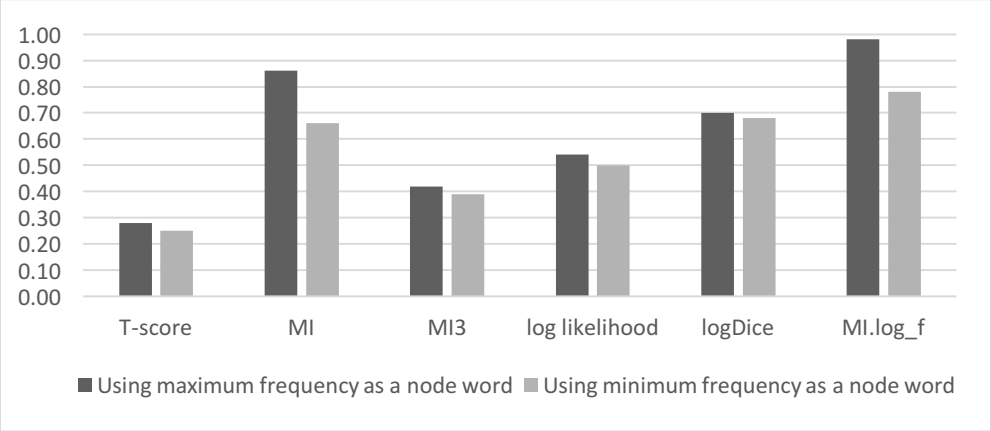


Figure 5.10: Comparing the MAP scores for the two datasets.

Figure 5.11 summarises the results of the AM evaluation of the two data sets by calculating the average MAP scores for both. This shows that the MI.log_f and MI were ranked as the best AMs for predicting the right collocates of the Arabic keyword list. This result is in line with Alrabiah et al. (2014) and another extensive empirical evaluation of 87 AMs in the automatic extraction of Czech collocations by Pecina (2005), who found that Pointwise MI measures achieved the best result with a 73.0% precision score.

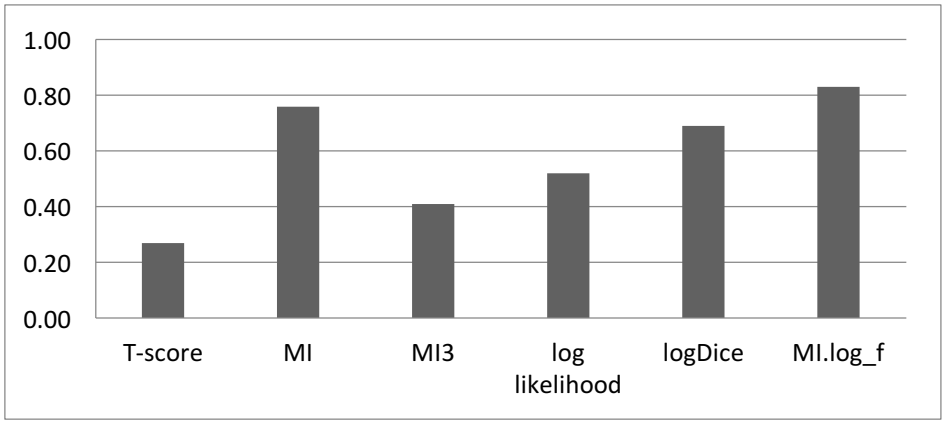


Figure 5.11: The average MAP scores for both data sets.

Table 5.17 presents an example of the MWEs extracted. It shows that these bigrams represent various syntactic constructions and semantic fields as the current study was not restricted to syntactic structures or the semantic domain.

Table 5.17: Examples of extracted MWEs with their syntactic structures.

MWEs	Structures
من أجل min ʾajl *'in order to'*	Prep-Noun
اعتماداً على iʿtimādan ʿalā *'based on'*	Noun-Prep
التنافس المحموم attanāfus almaḥmūm *'frenzied competition.'*	Noun-Adj
مدرج المطار madraj almaṭār *'airport Runway.'*	Noun-Noun
ظروف قاسية ḏurūf qāsiya *'severe conditions.'*	Noun-Adj

Figure 5.10 presents a comparison between the findings of the two rounds of the experiment. A slight drop can be noted in the performance of all AMs, as can a change in the ranking of the best AMs in that the MI achieved the second-best AMs when using less common node words. The t-score is still the least accurate AM in terms of predicting MWEs, regardless of the level of frequency of the node words.

5.7.6 Summary

Thus far, a brief report has been presented on an empirical study that aimed to evaluate the best AMs in the process of extracting AMWEs. This work is part of a series of experiments that use a statistical and symbolic approach to retrieve various types of semantically regular and high-frequency MWE in order to build intensive AMWE LRs for use in LP and NLP. The evaluation of AMs in this study shows a superior predictive result for AMs when using high-frequency data. The MI.log_f, MI and logDice achieved the highest precision scores in the extraction of MWEs from large SA corpora. Thus, these AMs are the best candidates when it comes to predicting useful and meaningful MWEs related to a frequency based Arabic wordlist. On the other hand, the MAP scores illustrate that T-score and MI3 are the worst AM candidates in predicting useful MWEs, while the L.LK can be seen as a potentially useful candidate in extracting meaningful MWEs.

5.8 Comparison and error analysis

The finding of previous experiments on the comparative evaluation of AMs in ranking AMWE candidates based on various reference datasets provides informative insights into the task of selecting and evaluating several AMs. In the first three experiments, the objective was to compare the performance of AMs on three gold standard lists, and the findings generally show that AMs record the best overall performance on

nominal AMWE bigrams followed by verbal, prepositional and other kinds of AMWEs used in dataset 3.

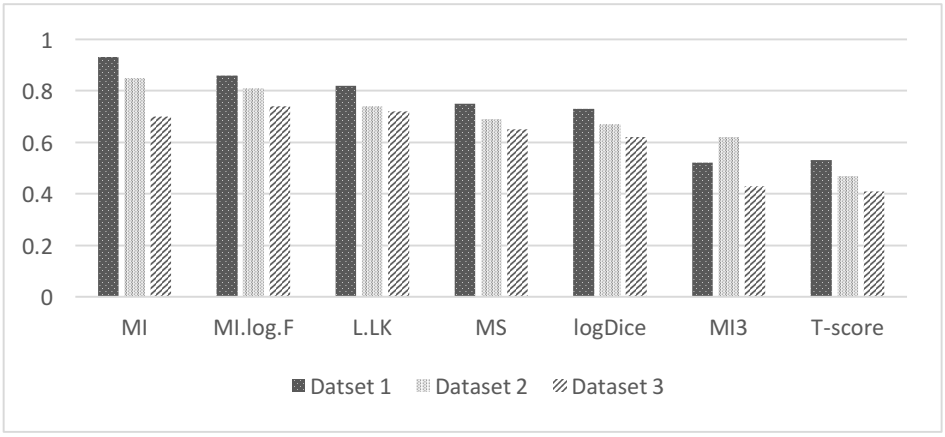


Figure 5.12: The MAP scores of AMs applied to the three datasets.

Figure 5.12 presents a summary of the averaged MAP scores for all AMs implemented in the first three experiments in this chapter. As expected, MI and MI.log.F were the overall best-performing AMs on the multiple evaluation datasets used in our evaluation experiments, followed by L.LK. MS and logDice were found to exhibit similar overall performance in the evaluation of around 0.65 while MI3 and T-score were the least predictive in ranking multiple types of AMs. However, the AMs are usually very sensitive to the kinds of data used in the evaluation tasks which makes it difficult to claim that these measures will always achieve the best result in extracting AMWEs. However, in a similar experimental setting and with similar data types these measures would be expected to achieve a similar result in ranking AMWE items. Furthermore, the evaluation tasks reported in this chapter can be replicated on various types of datasets or on a larger scale comparative evaluation in future work to achieve more reliable and informative results.

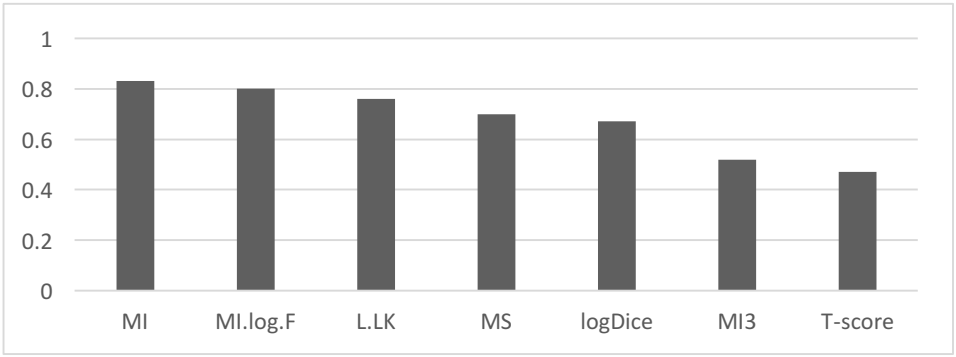


Figure 5.13: The overall precision scores of AMs applied to the three datasets.

In experiment 4, the aim was to compare the performance of AMs between high and low-frequency candidates and provide a method for extending and updating the reference lists by extracting the collocation of frequency-based word lists of SA. The finding reveals evidence of better performance in AMs when applied to high-frequency items in comparison to less frequent items. This is one of the limitations of AMs that should be taken into consideration when implementing these methods on low-frequency data sets.

To examine whether the performances of these AMs on various datasets is statistically distinct, the Student's t-test was applied to explore the differences between the experimental findings. The test results found only one statistically discernible difference with a significance score of < 0.05 between the performances of AMs in experiments one and two, which were described in sections 5.4 and 5.6 in this chapter. The full results for the significance tests are presented in Table 5.18. However, the size of the experiments and the nature of the reference data play a significant role in interpreting these results.

Table 5.18: The result of the significance tests (Student's t-test).

Pairs	Datasets	Sig. (2-tailed)
Pair 1	Dataset 1 & Dataset 2	.134
Pair 2	Dataset 1 & Dataset 3	.000
Pair 3	Dataset 2 & Dataset 3	.013
Pair 4	Dataset 4.A & Dataset 4.B	.061

The quantitative analysis of AMWE candidates classified as false positives in the evaluation experiments shows various types of errors in the extracted candidates that will be described briefly with examples from the findings of the previous four experiments. Table 5.19 presents examples of invalid AMWE candidates derived from the n-best lists in the AM evaluation tasks. The erroneous instances can be classified into various classes based on the type of error; for instance, several false items were found because of the error in automatic linguistic tokenisation and annotation, as can be seen in the two expressions*, faḍṭr ʾilā* and *dā atḫḏ*. Another error was attributed to inadequate lemmatisation which leads to data redundancy, as can be seen in the candidate baʾiḍāfthā which is a variant of the AMWE *bālāḍāfa*. Other errors were caused by the limited coverage of the reference list used in the

evaluation task, as shown in the expression *manqṭ ʿ annaḏīr* which is a valid AMWE but does not exist in the evaluation list. A further type of error was related to the excluded categories of AMWE in the study such as NEs or terminological terms.

Table 5.19: Samples from false AMWE candidates along with types of error.

AMWE candidates		Type of error
في على	fī ʿalā	non-AMWEs
فاضطر إلى	faḍṭṭur ʾilā	linguistic annotation
بإضافتها	biʾiḍāfatihā	morphosyntactic variation
حسني مبارك	ḥusnī mubārak	NEs
دا اتخذ	dā ittaḫaḏ	tokenisation
دقة تيموثي	diqqa taymūṯī	spelling error
منقطع النظير	munqaṭiʿ annaḏīr	not found in the reference dataset

5.9 Summary and Conclusion

Thus far in this chapter, several comparative evaluations have been presented which measured the performances of several well-known AMs on multiple types of previously developed reference datasets. Following similar procedures and experimental settings, four main AM evaluation tasks were implemented. The findings show an advantage for using three AMs which yields the most insightful and predictive result in the ranking of AMWE candidates based on multiple bigram reference lists. The recommendation is therefore to adopt these AMs in the AMWE extracting task implemented in a similar framework and in the context of the current research. The evaluation experiments performed here can also be reapplied using the same procedures on various AMWE evaluation datasets to examine potential similarities and differences. Furthermore, the current study can be extended to a larger scale evaluation task given the availability of comprehensive gold standard AMWE lists that represent a wide range of MWEs in SA.

6 Automatic Extraction of AMWEs Based on Morphosyntactic patterns and Association Measures

6.1 Introduction

Based on the findings of the AMWE extraction experiments reported in chapters 4 and 5, this chapter reports an additional four extraction experiments involving the implementation of an automatic extraction model for AMWE discovery from a large annotated SA corpus. The discovery model was mainly adopted from the previous study reported in section 4.5 with several modifications that extend the morphosyntactic patterns used and also benefits from the results of the AM evaluation studies reported in chapter 5, as will be illustrated in this chapter when relevant. The results of these experiments will be used to extend and update the AMWE reference lists and enhance the AMWE lexicon developed in this thesis.

The experimental findings experiments were quantitatively evaluated by manual and automatic annotation of the output against previously constructed gold standard lists of AMWEs. The annotated ArTenTen corpus was used in the AMWE extraction study. The use of linguistic annotation in AMWEs extraction is ideally the most appropriate solution for eliminating noisy data and concentrating the extraction task on the most valuable lexical units. Several studies have highlighted the significant impact of linguistic metadata in the improvement of MWE extraction and identification tasks (e.g., Smadja et al. 1996; Pearce 2002; Krenn et al. 2004; Evert 2004).

One of the main obstacles for NLP research progress in SA and other LR languages is the lack of publicly available and well developed LRs with a rich linguistic annotation, which makes it a challenging task to implement MWE extraction experiments based on richly annotated corpora. Thus, in the extraction process, linguistic annotation is applied to the corpus by using the available SA toolkits for morphology and shallow syntactic disambiguation to enhance the final output of the extraction model.

This chapter is organised as follows. Section 6.2 describes the AMWE extraction model, and the three AMWE extraction experiments are reported in sections 6.3 to 6.5. Section 6.6 presents the evaluation and validation findings of the extraction experiments. In section 6.7 and 6.8 the primary results are discussed along with several examples of erroneous candidates found in the evaluation tasks. Finally, a summary and conclusion are presented in section 6.10.

6.2 Method: AMWE Extraction model

The AMWE extraction model consists of a series of phases which ultimately result in the automatic extraction of several lists of AMWEs based on multiple sets of selection syntactic patterns in SA from a large annotated corpus. Figure 6.1 shows the main stages of the AMWE discovery model implemented in the extraction experiments based on various morphosyntactic patterns.

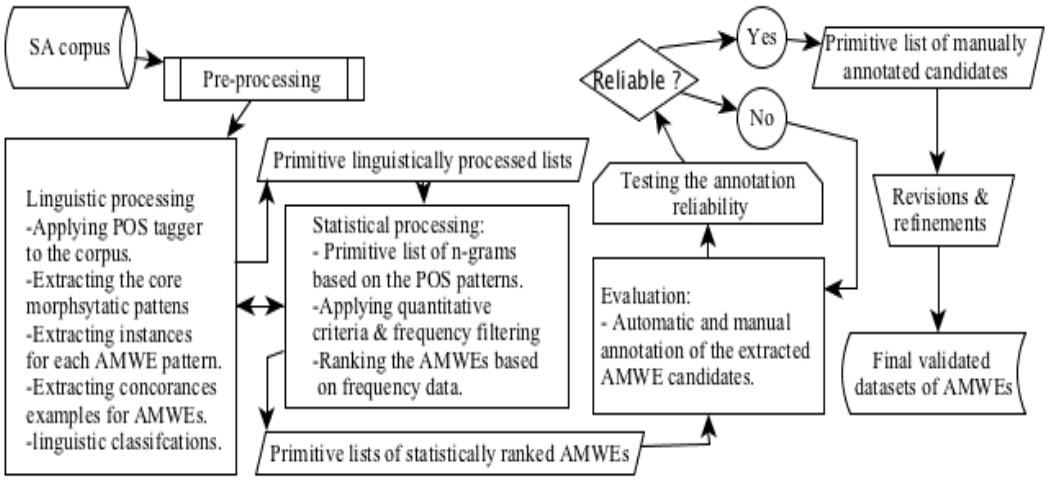


Figure 6.1 Diagram of the hybrid extraction model based on multiple AMWE morphosyntactic patterns.

As mentioned previously, the hybrid extraction model implemented in the development of the reference list illustrated in section 4.5 was adopted. The extraction processes implemented on the ArTenTen corpus, more information about the corpus, and the reasons for using it have previously been described in section 4.3.

The extraction model primarily consists of three main stages which include several types of text processing and extraction subtasks including linguistic, statistical, evaluation and validation phases. These main phases were generally conducted in a

sequential order although there was an occasional overlap between several linguistic and statistical processing subtasks to ensure the best possible output results for the extraction model were obtained. The evaluation of experiments in this chapter is based on automatic and manual annotation of the extracted AMWE candidates. In the automatic annotation, the candidates were matched against gold standard reference lists of AMWEs developed in previous work where well-validated reference lists of multiple types of AMWEs were constructed for use in various MWE extraction experiments, as illustrated in chapter 4 of this thesis.

The first phase prepares the corpus by removing any duplications and making the usual normalisation tasks to reduce noisy data, followed by the automatic morphological analysis which includes several linguistic analysis tasks such as tokenisation, lemmatisation, and POS tagging. However, to achieve the best possible output at this stage of the model in relation to the automated linguistic analysis, both SAP and MA linguistic toolkits were used which are two well developed and evaluated toolkits that have been used intensively in various ANLP tasks. More details about these tools and the linguistic processing is provided in sections 4.4 and 4.6.2.

In the statistical stage, the n-grams statistical model was used to extract the morphosyntactic selection patterns that represent various constructions from 2 to 6 components based on the linguistic annotation applied in the previous stage. Moreover, several AMs were used to extract multiple bigrams based on the findings of the comparative AM evaluation reported in chapter 5 which presents a favourable performance for MI, L.L.K and MI.L_F AMs in extracting AMWEs. The retrieved bigram was also used in a post-processing phase to retrieve longer AMWE candidates by joining the bigrams with other lexical units that have a strong affinity. This follows previous studies that have used this method of extending the extracted bigram (e.g., Kim et al., 2001; Seretan, 2011).

At this stage, for each experiment reported in this chapter selection morphosyntactic patterns will be used that were found in earlier studies and other related work described in section 4.6.3. Furthermore, a list of more complex morphosyntactic patterns was added that includes frequent low candidates of AMWEs to extend the coverage of the discovery model.

In the next step, the selection patterns were used to extract multiple types of AMWEs from the corpus based on the targeted AMWE constructions. These focus on nominal expressions in the first experiment, verbal expressions in the second experiment, and prepositional and other types of AMWE constructions in the third experiment. Thus, the model implemented in these various experiments resulted in the extraction of several AMWE lists which were then evaluated by manual and automatic annotation using the reference lists of AMWEs that were developed and manual annotation for newly discovered AMWEs.

To sum up, the following procedures have been conducted mostly in sequential order but with several overlaps between many extraction stages to arrive at the best possible findings and discover the most useful AMWE items to enhance the reference lists and the developed lexicon for AMWEs. The primary extraction steps were as follows:

Preprocessing and corpus preparation tasks.

Automatic linguistic analysis and POS annotation by SAP.

Selecting the most predictive morphosyntactic patterns for discovering AMWE candidates.

Using statistical techniques, specifically the n-grams and AM models, for extracting multiple types of AMWE Candidates.

Post-processing stage which aims to enhance the retrieved bigrams with other related items to cover longer and complicated candidates.

Candidate filtering, based on statistical data and linguistic annotation criteria, where the candidate list was reduced to a manageable and feasible number for evaluation purposes.

Error analysis which aims to determine the main obstacles and problems that prevent the extraction model from generating the best possible output.

Manual and automatic evaluation by aligning the extracted items to the previously developed reference lists of AMWEs or by manual annotation and classification in the event of limited coverage of the reference lists.

Final refinements and revisions of the extracted AMWE items which prepare the data sets for inclusion in the AMWE lexicon.

However, this is by no means an exhaustive list of all the tasks conducted, but it does focus on the main steps which might include several other processing subtasks. Nevertheless, more details about the extraction experiments are provided in the following subsections in which three reports are presented on the AMWE extraction experiments conducted.

6.3 The extraction of nominal expressions

In this section, an extraction experiment is reported that aimed to discover multiple types of nominal AMWEs based on several morphosyntactic selection patterns. As stated in many MWE studies in English, Arabic, and other modern languages, nominal MWE appears to be the most common and dominant type of MWE found in the literature (e.g., Najar et al., 2016; Meghawry et al., 2015b; Attia et al., 2010a; Vincze et al., 2011a; Castagnoli et al., 2014). Hence, in this study, the aim was to focus more intensely on this phenomenon in AMWE by using a hybrid model to discover various nominal AMWEs based on the most predictive morphosyntactic selection patterns. The research conducted in this chapter benefited considerably from the previous AMWE experiment and evaluation reported in chapters 4 and 5 of this thesis and also the work published in Alghamdi and Atwell (2017) and Alghamdi and Atwell (2016a).

The extraction of nominal AMWEs in this chapter was based on the hybrid model with several modifications described in detail in section 4.5 and in section 6.2. In this experiment, several extraction patterns were applied for use in the linguistic part of the hybrid extraction model. As mentioned in section 4.6.3, three primary sources were used to pick the most predictive extraction patterns. The SAP toolkit used in the automatic linguistic tasks in this study has multiple tags which represent various types of the noun in SA. Table 6.1 list the nominal tags used with examples.

Table 6.1: Nominal tagset used by SAP in the POS tagging.

tags	description	example	
NN	noun, singular or mass	شروط القيد	šurūṭ alqayd
DTNN	noun, singular or mass with the determiner "Al" (ال)	الادوية الموضعية	al'adwiya almuwḍiʿyya
NNP	Proper noun, singular	الله أكبر	ʾallāhu ʾakbar
DTNNS	noun, plural with the determiner "Al" (ال)	المعلومات الشخصية	almaʿlūmāt aššaḫṣiyya
NNS	noun, plural	قوات التحالف	quwwāt attaḥāluf
DTNNP	Proper noun, singular with the determiner "Al" (ال)	القرآن الكريم	alqurʾān alkarīm
DTNNPS	Proper noun, plural with the determiner "Al" (ال)	الآيات الكونية	alʾāyāt alkawniyya
NNPS	Proper noun, plural	خلوات روحية	ḫalwāt rūḥiyya

The statistical n-gram model and linguistic annotation were then used to generate several frequency-based nominal selection patterns lists which provide an overall view of the most frequent morphosyntactic nominal patterns found in the corpus. The extracted lists represent various expression lengths from 2 to 6 components. Table 6.2 presents examples from the most frequent morphosyntactic patterns and shows that the noun class was dominant among these POS patterns.

Table 6.2: Examples of patterns discovered for nominal AMWEs.

N-gram	pattern	N-gram	pattern
2	NN PRP$	4	NN PRP$ NN DTNN
	NN DTNN		NN DTNN CC DTNN
	NN NN		DTNN NN PRP$ NN
	NN NNP		NN PRP$ NN NN
	NN IN		DTNN NN PRP$ DTNN
	NNP NN		NN DTNN IN NN
	NN JJ		NN DTNN CC NN
	NNP NNP		DTNN IN NN DTNN
	NNP DTNN		DTNN IN NN NN
	NNP IN		NN NN NN DTNN

	NN PRP$ NN		DTNN CC DTNN CC DTNN
	NN NN DTNN		DTNN NN PRP$ NN DTNN
	NN PRP$ DTNN		NN DTNN NN PRP$ NN
	DTNN CC DTNN		NN DTNN CC NN DTNN
3	DTNN IN NN	5	NN DTNN IN NN NN
	DTNN NN PRP$		NN DTNN PUNC CC VBD
	NN NN NN		NN NN DTNN CC DTNN
	NN DTNN DTJJ		DTNN DTJJ IN NN DTNN
	NN IN NN		NN DTNN IN NN DTNN
	NN DTNN NN		NN DTNN DTJJ IN NN
	NN DTNN CC DTNN CC DTNN		
	NNP IN NN NNP IN NN		
	NN DTNN NN PRP$ NN DTNN		
	NN PRP$ NN DTNN CC DTNN		
6	NN NN PRP$ CC NN PRP$		
	NN DTNN DTJJ IN NN DTNN		
	NN PRP$ NN PRP$ NN DTNN		
	NN NN DTNN NN PRP$ NN		
	NN DTNN CC DTNN NN PRP$		
	NN DTNN DTJJ NN PRP$ NN		

These morphosyntactic patterns were then used for several corpus tests to generate various candidate lists which then underwent manual quantitative analysis to select the most appropriate patterns. Due to the limited scale and restrictions of the current research, the extraction was limited to only 12 selection patterns. Table 6.3 provides examples of the patterns used in this experiment which encompass various types of common pattern that represent the multiple constructs of AMWEs.

Table 6.3:Examples of selection patterns used in the extraction of nominal AMWEs.

Pattern	MWE candidate	Pattern	MWE candidate
NN, DTNN	توفير الحصانة	DTNNP-IN-DTNN	الناي ب النفس
	سجناء الحرية		النهار ب الليل
	عقلية التشكك		المنزه عن الخطأ
NN-NN	خير مثال	NN-DTNN-CC-DTNN	عبر الصحافة و الاعلام
	ازمان مقبل		طب الكوارث و الطوارئ
	وراء اسوار		سجناء الحرية و العدالة
DTNN, DTJJ	الرخاء الاقتصادي	NN-DTNN-IN-NN	عدم الانجرار الى تقييد
	الشرائح المجتمعية		خلال الوصول الى نقطة
	الصعيد الدولي		درجة التساؤل عن جدوى مع الاخذ ب عين
NNS-IN-NN	وجهان لعملة	NN-DTNN-CC-NN - DTNN	سداد الخطى و بلاغ المنى
	خطوات ب اتجاه		ذات اليمين و ذات الشمال
	عصفورين ب حجر		محو الامية و تعليم الكبار
NN-IN-NN	متحدث ب اسم	DTNN-CC-DTNN-CC-DTNN	التسلط و القهر و الظلم
	يوما بعد يوم		النزاهة و الاخلاص و التفانى
	جنبا الى جنب		القحط و الجفاف و المجاعة

As shown, the selection patterns at this stage of processing includes various AMWE structures from simple two token compound phrases to longer items with five components in contiguous and non-contiguous candidates.

As expected, the short selection patterns that include two or three components yield the most predictive results in the extraction process. For instance, the pattern [N-N] is one of the most common compound nouns of AMWEs. Under this main pattern, several variants of nominal structures can be found which cover different types of syntactic relation (e.g., [NN-DTNN]- [NN NN]- [DTNN NN] - [DTNN DTJJ]). Because of the limited scale of the current experiment, several restrictions were imposed in the extraction process which included limiting the extraction patterns used to 12 patterns with a threshold frequency of 10 per million words in the candidate filtering stage.

The extraction model involves several linguistic processing tasks that have been conducted in the extraction process, beginning with the pre-processing and preparation tasks such as the normalisation of SA script described in section 2.2.1.1. Other tasks relate to automatic linguistic disambiguation, which includes the typical

linguistic processing pipeline (e.g., tokenisation, lemmatisation, and morphological and syntactic annotation). The linguistic information and shallow syntactic analysis were also used to extract the most predictive selection morphosyntactic patterns which is an essential part of the extraction model. The statistical processing tasks in the extraction model involve corpus indexing and the use of an n-gram model to generate multiple AMWE lists using various selection patterns. Furthermore, based on the evaluation findings reported previously in chapter 5 of this thesis, the best AM predictors were used to sort the bigram generated candidate lists in descending order, according to which AMs assist in the process of filtering out undesirable instances generated by the extraction model. Table 6.4 presents samples of bigram AMWE candidates sorted according to the MI and MI.log.F AMs.

Table 6.4: Samples of bigram AMWE candidates sorted by MI and MI.log.F AMs[54]

AMWE bigram		MI score	AMWE bigram		MI.log.F score
المتحدة الانمائي	almuttaḥida al'inmā'ī	5.42865	التعاون الخليجي	atta'āwun alḫalījī	36.547
المجلس الانتقالي	almajlis al'intiqālī	5.41233	المجلس الوطني	ilmajlis alwaṭanī	34.725
وزيرة الخارجية	wazīrat alḫārijyya	5.05357	المقاومة الإسلامية	almuqāwama al'islāmiyya	33.284
هيلاري كلينتون	hīlārī kilīntūn	4.98237	البريد الالكتروني	albarīd alilikturūnī	33.076
منظمة التحرير	munaḍḍamat attaḥrīr	4.82744	السبت الموافق	assabt almuwāfiq	32.621
ميثاق الامم	miyṯāq aluumam	4.43827	الخارجية السوري	alḫārijyya assūwrī	32.604
تردي الأوضاع	taraddī al'awḍā'	4.41279	التحرير الفلسطينية	attaḥrīr alfilasṭīniyya	32.407
صندوق النقد	ṣundūq annaqd	4.37249	التعليم العالي	atta'līm al'ālī	32.314
جبهة التحرير	jabhat attaḥrīr	4.33835	الدول العربية	adduwal al'arabiyya	32.217
امتحانات الثانوية	imtiḥānāt aṯṯānawiyya	4.24463	الوحدة الوطنية	alwaḥda alwaṭaniyya	32.211

As seen in these examples, the use of AMs as a type of statistical filtering eliminates the extraction of a considerable number of irrelevant items. The bigram extracted in this study was used to retrieve longer candidates in a post-processing phase, for instance, the two bigrams *ʿalā arraġm* and *man ʾajl* can be extended by adding

[54] [DTNN-DTJJ] and [NN-DTNN] morphosyntactic patterns used in the presented AMWE examples.

multiple words that have strong affinities with them such as *ʿalā arraġm man ʾan* and *man ʾajl ʾan*. Furthermore, the model permits the extraction of non-contiguous candidates using various regular expression tactics, as described in section 4.5.2. The retrieved nominal AMWEs represent a wide range of syntactic structures and multiple length types, ranging from bigram MWE candidates to other expressions that consist of 5 or more components. However, this does not include non-contiguous expressions with word interventions that were allowed in the extraction model and which might involve slots that comprise one to four components. The regular expression[55] functions were used to extract non-contiguous AMWE and match various morphological variations of the retrieved items. Examples of these regular expressions in Python language formalism are provided in Table 6.5.

Table 6.5: Common regular expressions in Python language.[56]

Special characters	Function
Dot.	In the default mode, this matches any character except a new line. If the DOTALL flag has been specified, this matches any character including a new line.
Caret.	Matches the start of the string, and in MULTILINE mode also matches immediately after each new line.
$	Matches the end of the string or just before the new line at the end of the string, and in MULTILINE mode also matches before a new line… etc.
*	Causes the resulting RE to match 0 or more repetitions of the preceding RE, as many repetitions are possible. ab* will match ‘a’, ‘ab’, or ‘a’ followed by any number of ‘b’s.
?	Causes the resulting RE to match 0 or 1 repetitions of the preceding RE. ‘ab’ will match either ‘a’ or ‘ab’.

An example of discontinuous AMWE can be seen in Table 6.6 which shows AMWEs that primarily consist of 3 core components with multiple types of intervening words

[55] The notion of regular expression or what is known as (regex) or (regexp) dates back to the 1970s and can be defined as ‘An expression that describes a set of strings (= a regular language) or a set of ordered pairs of strings (= a regular relation). A finite-state automaton can represent every language or relation described by a regular expression. There are many regular expression formalisms. The most common operators are concatenation, union, intersection, complement (=negation), iteration and composition. Also called rational expression.' (Mitkov, 2005, p. 706).

[56] Examples and description of regular expressions from the python online documentation. For more details see https://docs.python.org/3/library/re.html#re-syntax.

between the main components of the expression. Figure 6.2 shows a pattern of regular expression used to match one type of discontinuous AMWE construct in this study.

```
(1:[tag="(DT)?NN.*" | tag="PRP.?" ] 2:[] 3:[] 4:[] 5:[tag="(DT)?JJ.*"] & f(2.tag)
```

Figure 6.2: Regular expression patterns for the structure N-A within a gap of 3 tokens.

Different types of intervening words were found in this experiment. As shown in Table 6.6, the intervening slots in flexible AMWE candidates range from one token to more than four contiguous tokens. The pronoun also seems to be one of the most common word classes in the intervening words. However, this is an anticipated finding because pronouns are used as joint words in most types of SA nominal sentence. Other types of flexible AMWEs, which include nested items within the intervening words, were excluded in the study because processing these types of lexical units requires manual processing which is a time-consuming task for a PhD project.

Table 6.6: Example of multiple intervening words in a nominal AMWE candidate.

last part	Intervening words		Initial part
لعملة	مختلفين و لكن هما	muḫtalfayn wa lākin humā	wajhān وجهان
liʿumlat	ل حدثين تاريخيين	li ḥadaṯayn tārīḫiyyayn	
	متلازمان	mutalāzimān	
	غير مختلفان	ġayr muḫtalifān	
	ساطعان	sāṭiʿān	

The final finding of this experiment is that there are various initial lists of AMWE candidates which reflect multiple main morphosyntactic patterns selected in the linguistic processing phase with a total of 37.671 items. The candidate list then underwent several filtering processes which includes sorting the extracted bigram items according to best AMs and applying a frequency threshold to the lists based on the length of extracted items. Other filtering tasks implemented on the datasets include automatic identification of NEs and statistical and linguistic filtering as described in section 4.6.6. After the filtering phase, a total of 14.572 AMWE candidates remained which will be partially used in the evaluation task reported in section 6.6.

Table 6.7 presents samples of nominal AMWE candidates generated in this experiment; as shown, these candidates contain a variety of lexical units that represent multiple morphosyntactic and semantic variants. The final extracted items in this study were classified into 19 categories according to their morphosyntactic structures and the number of components in the expressions.

Table 6.7: Sample of the extracted lists of nominal AMWE candidates.

Structure		Examples
DTNN-DTNN	الدول العربية	adduwal alʿarabiyya
	المجتمع المدني	almujtamaʿ almadanī
	الشعب الفلسطيني	aššaʿb alfilasṭīnī
NN DTNN	كرة القدم	kurat alqadam
	اهل البيت	ʾahl albayt
	حقوق الانسان	ḥuqūq alʾinsān
DTNN-DTJJ	الاجهزة الامنية	alʾajhiza alʾamniyya
	الآونة الاخيرة	alʾāwina alʾaẖīra
	الاتحاد الاوروبي	alʾittiḥād alʾūrūbbī
DTNNS-DTJJ	المؤسسات الرسمية	almuʾassasāt arrasmiyyat
	التأمينات الاجتماعية	attaʾmīnāt alijtimāʿiyya
	الاجراءات الجزائية	aliʾijrāʾāt aljazāʾiyya
DTNN-IN	التعامل مع	attaʿāmul maʿa
	المشاركة في	almušāraka fī
	العلاقة بين	alʿalāqa bayn
NN-IN-NN	عبارة عن مجموعة	ʿibāra ʿan majmūʿa
	متحدث باسم	mutaḥaddiṯ bism
	تصريح لوكالة	taṣrīḥ liwikāla
NN-DTNN-CC-DTNN	اهل السنة والجماعة	ʾahl assunna waljamāʿa
	بين الحين والاخر	bayn alḥīn wālaʾāẖar
	حرية الراي والتعبير	ḥurriyyat arraʾy wa attaʿbīr
NN DTNN CC DTNN CC DTNN	ذي القربى و اليتامى و المساكين	ḏī alqurbā wa alyatāmā wa almasākīn
	معادلة الجيش و الشعب و المقاومة	muʿādalat aljayš wa aššaʿb walmuqāwama
	معاهدة الاخوة و التعاون و التنسيق	muʿāhadat alʾuẖuwwawa attaʿāwun wa attansīq

In the next step, random samples from the retrieved candidates will undergo final evaluation by manual and automatic annotation as will be described in section 6.6 of this chapter.

6.4 The extraction of verbal expressions

Verbal AMWEs are found in a significant proportion of SA and also in other languages as reported in recent MWE research that focused on verbal MWEs (e.g., (Bejcek et al., 2017; Todiraşcu et al., 2008; Taslimipoor et al., 2012). Further details are provided in section 2.3 of this thesis. As described in section 3.2.1.4, the SA sentences were divided into two main categories: nominal and verbal sentences. The second type includes phrases that start with various verb types such as past, imperative, and contiguous. Before describing the experiment conducted in this study, a brief description will be presented of the linguistic properties of verbs and verbal constructs in SA that assist in understanding the results obtained in the current study on verbal AMWEs. However, an in-depth linguistic illustration of the verb system in SA can be found in several research studies (e.g., Badawi et al., 2013; Ryding, 2005).

As a Semitic language, SA is morphologically rich and this is evident in the morphological behaviours of verbs which exhibit several variations based on various linguistic functions. Hence, the verbs in SA have many conjugations that are marked by common grammatical categories which include stem, person, number, tense, gender, mood, and voice. Table 6.8 presents the main grammatical features of verbs with examples. There are 26 core verb patterns in SA which include six trilateral and one basic quadrilateral pattern in addition to 19 augmented forms, as shown with examples in Table 6.8. These types of verb pattern are also summarised in Figure 6.7 which presents a hierarchy of the main morphological classes of verbs in SA.

Table 6.8: Grammatical categories of SA verbs with AMWE examples.

Features	Values			Examples
Stem	basic	trilateral	كَتَب	kataba
		quadrilateral	وَسْوَس	waswas
	augmented	Includes a set of derived patterns (see table 6.4)	تَكْتُب	taktub
Aspect	perfect		كَتَبَ	kataba
	imperfect		يَكتب	yaktub
	imperative		اكتُب	ʾuktub
Voice	active		كَتَبَ	kataba
	Passive		كُتِبَ	kutiba
Person	1st		اكْتُب	ʾaktub
	2nd		تَكْتُب	taktub
	3rd		يَكتُب	yaktub
Gender	masculine		يَكتُب	yaktub
	feminine		تَكتُب	taktub
Number	singular		كَتَبَ	kataba
	dual		كَتَبَا	katabā
	plural		كَتَبُوا	katabū
Mood	indicative		يكتبُ	yaktubu
	subjunctive		لَن يَكتُبَ	lan yaktuba

Table 6.9: Core basic and augmented verb patterns in SA.

Stem	Patterns		Examples	
basic	فَعَل يَفْعُل	faʿal yafʿul	كتب يكتب	katab yaktub
	فَعَل يَفْعِل	faʿal yafʿil	كسر يكسر	kasar yaksir
	فعَل يَفْعَل	faʿal yafʿal	ذهب يذهب	ḏahab yaḏhab
	فَعِل يفعَل	faʿil yafʿal	شرِب يشرب	šarib yašrab
	فَعُل يفْعُل	faʿul yafʿul	حسُن يحسن	ḥasun yaḥsun
	فعِل يفْعِل	faʿil yafʿil	حسب يحسب	ḥasib yaḥsib
	فَعْلَل يُفَعْلِل	faʿlal yufaʿlil	دحرج يدحرج	daḥraj yudaḥrij

	أَفْعَلَ	ʾafʿal	أنزل	ʾanzal
augmented	فَعَّل	faʿʿal	كسَّر	kassar
	فاعل	faʿal	حاور	ḥāwar
	انْفَعَل	infaʿal	انكسر	inkasar
	افْتَعَل	iftaʿal	اجتمع	ijtamaʿ
	افْعَلَّ	ifʿall	اخضر	iḫḍarr
	تفعَّلَ	tafaʿʿal	تعلم	taʿallam
	تفاعل	tafāʿal	تحاكم	taḥākam
	استفعل	istafʿal	استغفر	istaġfar
	افعوعل	ifʿawʿal	اعشوشب	iʿšawšab
	افْعَوَّل	ifʿawwal	اجلوذ	ijlawwaḏ
	افعالَّ	ifʿāll	اخضار	iḫḍārr
	تَمَفْعَل	tamafʿal	تمسكن	tamaskan
	تَفَوْعَل	tafawʿal	تجورب	tajawrab
	تَفَيْعَل	tafayʿal	تسيطر	tasayṭar
	تَفَعْلَل	tafaʿlal	تجلبب	tajalbab
	تَفَعْيَلَ	tafaʿyal	ترهيأ	tarahyaʾ
	تَفَعْلَى	tafaʿlā	تسلقى	tasalqā
	افْعَنْلَلَ	ifʿanlal	احرنجم	iḥranjam
	افْعَلَلَّ	ifʿalall	اطمأن	iṭmaʾann
	افْعَنْلَل	ifʿanlal	اقعنسس	iqʿansas
	افْعَنْلَى	ifʿanlā	احزنبى	iḥzanbā
	افْتَعْلَى	iftaʿlā	استلقى	istalqā

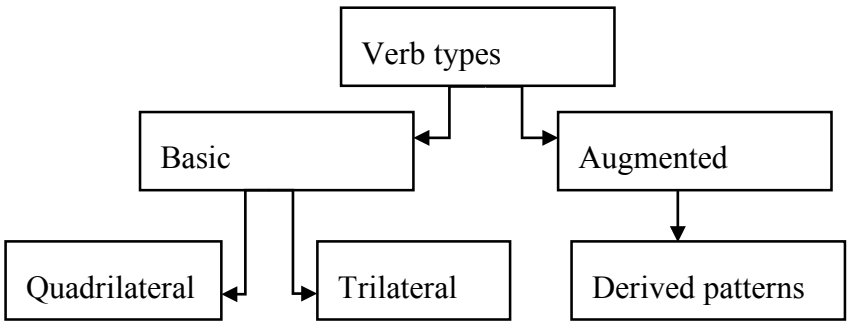


Figure 6.3: The core verb forms in SA.

Regarding verb behaviour in SA sentences, several essential points should be considered during the extraction process which summarises the varieties of verb forms in a different context. One of the primary rules in the verbal structure is that the verbs are usually modified by the subject of the phrases; thus, prefixes and suffixes

generally change according to the type of subject (e.g., masculine, feminine, singular, dual, plural). Table 6.10 shows various forms of verbs according to the type of subject involved in the sentence.

Table 6.10: Examples of verbs modified by various subject types.

Singular subject			**Present verb form**	
أنا	ʾanā	I	أَرسُمُ	ʾarsumu
أنتَ	ʾanta	you (masc.)	تَرسُمُ	tarsumu
أنتِ	ʾanti	you (fem.)	تَرسُمِينَ	tarsumīna
هُوَ	huwa	he/it	يَرسُمُ	yarsumu
هِيَ	hiya	she/it	تَرسُمُ	tarsumu
Plural subject			Present verb form	
نَحنُ	naḥnu	we	نَرسُمُ	narsumu
أنتُم	ʾantum	you (masc.)	تَرسُمُونَ	tarsumūna
أنتُنَّ	ʾantunna	you (fem.)	تَرسُمنَ	tarsumna
هُم	hum	they (masc.)	يَرسُمُونَ	yarsumūna
هُنَّ	hunna	they (fem.)	يَرسُمنَ	yarsumna
Dual subject			Present verb form	
أنتُمَا	ʾantumā	you	تَرسُمَانِ	tarsumāni
هُما	humā	they (masc.)	يَرسُمانِ	yarsumāni
هُما	humā	they (fem.)	تَرسُمانِ	tarsumāni

Regarding the verb position in the sentence, SA generally allows for flexibility in word order in the sentence. However, verbs are not an exception in this case; thus, they can be found before or after the subject with no clear change in the meaning of the phrase, as shown in these two examples:

يكتب الطالب واجبه

yaktub aṭṭālib wājibah

الطالب يكتب واجبه

aṭṭālib yaktub wājibah. Tran. The student is writing his homework.

In addition to the variations mentioned above and derivations of SA verbs, several types of verbal noun can be derived from verbs by reconstructing the verb core roots into various templets and forms, as shown in the examples in Table 6.11.

Table 6.11: Examples of verbal noun patterns in SA.

Verbal noun patterns		Root verb	
فَعْل-رَسْم	faʿl-rasm	رَسَمَ	rasama
فُعُول-دخول	fuʿūl-duẖūl	دَخَلَ	daẖala
فَعال-ذهاب	faʿāl-ḏahāb	ذَهَبَ	ḏahaba
فِعالة-صناعة	fiʿāla-ṣināʿa	صَنَعَ	ṣanaʿa
فَعَل-أمل	faʿal-ʾamal	أَمِلَ	ʾamila

Regarding the linguistic processing of the corpus, in this study the SAP toolkit was used which has several tags for representing various types of verbs in SA. Table 6.12 lists the tags used in the POS annotation with AMWE examples.

Table 6.12: The SAP tagset of verb forms with AMWE examples.

Verb tag	Description	MWE example	
VBP	the third person singular present	بعرف الكل	biʿurf alkull
VBD	past tense	دام ظله	dām ḏilluh
VBN	past participle	تدر لبنا و عسلا	tudir laban wa ʿasal
VB	base form	اضف الى ذلك	aḍif ilā ḏālik
VBG	gerund or present participle	سرعان ما تلاشى	surʿān mā talāšā

The VBP tag is the most frequent form used in the POS annotation while the VBG tag was found least often in the data. Table 6.13 provides instances of the standard verbs in the corpus.

Table 6.13: List of the most frequent verbs in the corpus.

Verb		Frequency	Verb	Frequency
كان	kān	254560	يجب	yajib
قال	qāl	235179	يتم	yatimm
ليس	laysa	78179	جاء	jāʾ
صلى	ṣallā	70646	يعني	yaʿnī
سلم	sallam	63722	ذكر	ḏakar
يمكن	yumkin	62343	يقوم	yaqūm

In the extraction experiment, the same procedures were followed as described in section 6.2. Therefore, after pre-processing tasks and preparing the data for extraction, several lists of various types of verbal constructs were generated based on the n-gram model processing of linguistic meta-data. This step provides an overall picture of the

most frequent verbal morphosyntactic patterns in the corpus. Hence, several corpus-based extraction tests can be implemented for these selection patterns to produce a list of the patterns that can be used along with the previous selection of verbal patterns in the current AMWE extraction tasks. More details about previous selection patterns are presented in section 4.6.3. Table 6.14 provides lists of the selected verbal AMWE patterns that were among the top scores.

Table 6.14: Examples of interesting patterns discovered for verbal AMWEs.

N-gram	Pattern	N-gram	Pattern
2	VBP NN	4	VBP IN NN DTNN
	VBD NN		VBP IN NN NN
	VBP IN		VBP DTNN IN NN
	VBP DTNN		VBP NN NN DTNN
	VBP VBP		VBP PRP IN NN
	VBD DTNN		VBD IN NN DTNN
	VBD IN		VBD IN NN NN
	VBD NNP		VBD DTNN IN NN
	VBP NNP		VBP NN DTNN NN
	VBD VBP		VBD NN NN DTNN
3	VBP IN NN	5	VBP NN PRP$ NN DTNN
	VBP NN DTNN		VBD NN PRP$ NN DTNN
	VBP NN NN		VBP NN PRP$ NN NN
	VBD IN NN		VBP PRP IN WP VBP
	VBD NN DTNN		VBP NN PRP$ IN NN
	VBD NN NN		VBP PRP IN WP VBP
	VBD NNP IN		VBD NNP IN NN NNP
	VBP VBP NN		VBP IN PRP NN DTNN
	VBP DTNN NN		VBP IN PRP NN NN
	VBP DTNN IN		VBP IN PRP IN NN

6	VBD NNP IN PRP CC VBD
	VBD NNP IN PRP CC NN
	VBD NN NNP VBD NNP IN
	VBP NN PRP$ NN PRP$ DTNN
	VBP PRP IN PRP VBP NN
	VBP NN DTNN NN PRP$ DTNN
	VBP NN PRP$ NN NN DTNN
	VBP IN NN DTNN CC DTNN
	VBP DTNN NN PRP$ NN DTNN
	VBP PRP IN WP VBP NN

In the following step, based on previous findings and the corpus-based investigation of multiple verbal patterns, 12 patterns were selected to be used primarily in the extraction model. However, these patterns involve many variants which are also used in this study; the limited number of extraction patterns is justified by the limited scale and other constraints of the experiments. Table 6.15 presents multiple selection patterns with a list of AMWE instances. These patterns range from two to six component expressions and represent various verbal structures and semantic domains. With the use of multiple frequency thresholds based on the length of the selection patterns, the extraction model in this step generates lists of AMWE candidates comprising a total of 24.267 items that will undergo multiple candidate filtering in subsequent processing phases.

Table 6.15: Examples of selection patterns used in the extraction of nominal AMWEs.

Pattern	MWE candidate	Pattern	MWE candidate
VBD-DTNNS	تناول الوجبات كشفت التحقيقات تجاوز العقبات	VBP NN DTNN	يتواكب مع التوجه يكون مجرى الحديث يبدا سريان الحظر
VBP DTNN	تنقيح التراث تجوب البحار يتحقق الرضا	VBD IN NN DTNN	يؤدي الى اثارة الشك يوجه ب سرعة البت يفتقر الى لقمة العيش

VBP-IN	يتعلق ب تؤدي الى يعبر عن	VBD NN NN DTNN	غلب عليها طابع التحدي غدوا عبر دورة الزمان ترك بصمة مع الفريق
VBP-DTNN-NNP	يقتصر الامر على ينص القانون على يفتح الباب على	VBP NN DTNN CC DTNN	تصدر للاقراء و الافتاء تعجز اقلام و السنة يصعب قياسها و التنبؤ
VBP-IN-NNS	ترتبط بعلاقات يكيل بمكيالين يغني على ليلاه	VBP IN NN DTNN CC DTNN	يؤدي الى اثارة الشك و القلق تركن ل قوى الظلام و الضلال يقودها ل بر الامان و الوحدة

Furthermore, the extraction process permits the extraction of non-contiguous candidates by using functions of multiple regular expression to discover flexible verbal items, as can be seen in the examples provided in Table 6.16.

Table 6.16: Example of multiple intervening words in verbal AMWE candidates.

last part	intervening words		initial part
الطين بلة	هذا الوضع	hāḏā alwaḍʿ	يزيد
aṭṭiyn billat	بذلك	biḏālik	yazīd
	الامر علة و	alʾamr ʿilla wa	
	الشكوك و يزيد	aššukūk wa yazīd	
	النار اشتعالا و يزيد	annār ištiʿālā wa yazīd	

In the candidate filtering tasks in the statistical stage, the AMs were used to discover the most silent bigram candidates; Table 6.17 provides examples of the retrieved bigram listed in descending order based on MI and MI.L.F AMs.

Table 6.17: Samples of bigram AMWE candidates sorted by MI and MI.log.F AMs.[57]

AMWE bigram	MI score	AMWE bigram	MI.log.F score
وجد الباحثون	4.1819	شاء الله	31.32306
أينما كان	3.78983	يجعلني اعتقد	30.60497
رفع المتظاهرون	3.63965	يفرض علينا	29.75315
اكد المشاركون	3.43147	يمكن تصور	29.39925

[57] [DTNN-DTJJ] and [NN-DTNN] morphosyntactic patterns used in the AMWE examples presented.

اطيعوا الرسول	3.32011	ماذا يعني	29.3184
كان رجلا	3.24418	يكون المرء	28.92929
قال الحافظ	3.17968	اجد احدا	28.01563
راح ضحية	3.14661	تكاد تكون	27.68077
أشارت الصحيفة	3.08147	يقول موضحا	27.34263
صلى الله	3.07642	يرى البعض	25.88423

This task is essential in the removal of candidates considered a type of noisy data or irrelevant lexical units. Additional filtering of candidates was also applied to reduce the size of the final extracted list and focus the retrieval process on the most valuable AMWE candidates. The final lists of verbal AMWE in this study consisted of 13.287 candidates that will then be used in the evaluation task reported in section 6.6 of this chapter. Table 6.18 presents a list of extracted verbal AMWE candidates that represent the various morphosyntactic patterns used in this study.

Table 6.18: Sample of randomly selected verbal AMWE candidates.

Structure	Instances	
VBP-NNS	يصلي ركعتين	yuṣallī rakʿatayn
	تلبي احتياجات	tulabbī iḥtiyājāt
	يقدم خدمات	yuqaddim ẖadamāt
VBP-IN	تساهم في	tusāhim fī
	يتناسب مع	yatanāsab maʿ
	يجمع بين	yajmaʿ bayn
VBD-NN	تم إعداد	tamm ʾiʿdād
	كشفت مصادر	kašafat maṣādir
	اطلق سراح	aṭlaq sarāḥ
VBP-DTNN	يمكن القول	yumkin alqawl
	اقرأ المزيد	iqraʾ almazīd
	تجدر الاشارة	tajdur alʾišāra

VBD-DTNN	قال الامام	qāl al’imām
	اكد الدكتور	akkad adduktūr
	حان الوقت	ḥān alwaqt
VBP-NNP-NNP	يفرض رايه على	yafriḍ raʾyah ʿalā
	يجري على لسانه	yajrī ʿalā lisānih
	تضحك على نفسك	taḍḥak ʿalā nafsik
VBP-DTNN-IN	يتم العمل على	yatimm alʿamal ʿalā
	يقتصر الامر على	yaqtaṣir alʾamr ʿalā
	يفتح الباب على	yaftaḥ albāb ʿalā
VBP NN DTNN CC DTNN	تكون موضوع الدراسة و البحث	takūn mawḍūʿ addirāsa wa albaḥṯ
	يعود بين الحين و الحين	yaʿūd bayn alḥīn wa alḥīn
	يلبس ثياب النصح و الوعظ	yalbas ṯiyāb annuṣḥ wa alwaʿḏ

6.5 The extraction of prepositional and other types of AMWEs

In this experiment, the hybrid extraction model was used in the extraction of multiple types of prepositional AMWEs. Furthermore, in the extraction experiments other kinds of expressions such as adverbial and adjectival phrases were included on a smaller scale. This experiment is an extension of the previous experiment concerning the extraction of reference lists of AMWEs to build a comprehensive lexicon of AMWE that will help improve several NLP tasks. The initial findings on prepositional expressions in the previous studies reveals that these types of MWE are very frequent in SA. Thus, prepositional AMWEs will be the focus of this extraction experiment which aims to explore potentially new items and selection patterns.

Before reporting the current experiment, it is useful to illustrate briefly the linguistic properties of prepositional phrases in SA. One of the distinctive features of propositions is that they are uninflected and underived words, and the prepositional expressions consist primarily of a preposition followed by a nominal phrase and the head noun is always in the genitive or oblique case in a SA sentence.

Based on the automatic linguistic analysis of the SAP toolkit implemented in this study, the tag IN was used mainly to annotate most types of prepositions in the corpus. Table 6.19 shows a list of common propositions found in the data for this experiment with instances of AMWE candidates.

Table 6.19: Examples of particles annotated by IN tag with instances of AMWE.

Particle		AMWE candidates	
ل	la	للحصول على	lilḥuṣūl ʿalā
ب	ba	بكل بساطة	bikull basāṭa
الى	alā	إلى أبعد من ذلك	ʾilā ʾabʿad min ḏālik
عن	ʿan	عن التصدي لـ	ʿan attaṣaddī li
في	fī	في بعض الأحيان	fī baʿḍ alʾaḥyān
علي	ʿalī	على نطاق واسع	ʿalā niṭāq wāsiʿ
ان	an	ان شاء الله تعالى	in šāʾ ʾallāh taʿālā
ك	ka	ككذا وكذا	kakāḏā wakāḏā
من	man	من الجدير بالذكر	min aljadīr biḏḏikr
فيما	faymā	فيما يتعلق بـ	fīmā yataʿallaq bi
بينما	baynmā	بينما يرى البعض الآخر	baynamā yarā albaʿḍ alʾāḫar
كي	kay	كي يتسنى لـ	kay yatasannā li
منذ	manḏ	منذ وقت مبكر	munḏu waqt mubakkir
ريثما	rayṯmā	ريثما تهدأ الأمور	rayṯamā tahdaʾ alʾumūr
لما	lammā	لما مثل بين يدي	lammā maṯul bayn yaday

Two other tags (W?RB and CC) were also used by SAP in the POS annotation to indicate other types of particles included in the extraction model, as shown in Table 6.20 which shows multiple instances of the particles used in the study.

Table 6.20: Examples of particles tagged with W?RB and CC tags in SAP.

Tag	Particles Examples	
CC	و	wa
	ف	fa
	او	aw
	كما	kamā
	ثم	ṯumma
	لكن	lākin
	أم	ʾam
	بل	bal

W?RB	اما	ammā
	اذا	iḏā
	حيث	ḥayṯ
	قط	qaṭ
	كيف	kayf
	ربما	rubbamā
	لماذا	limāḏā
	كيف	kayf
	ربما	rubbamā

Thus, the prepositional expressions in this study encompass a wide range of expressions that begin with multiple types of particles[58] they include several adverbs, prepositionals,[59] and others. Figure 6.4 presents a hierarchy of the core types of particles in SA. Particles in these categories can also be classified into different categories such as bound and free classes or into three main classes based on the types of word that follow them which can be nouns, verbs, or shared particles that can proceed both nouns or verbs.

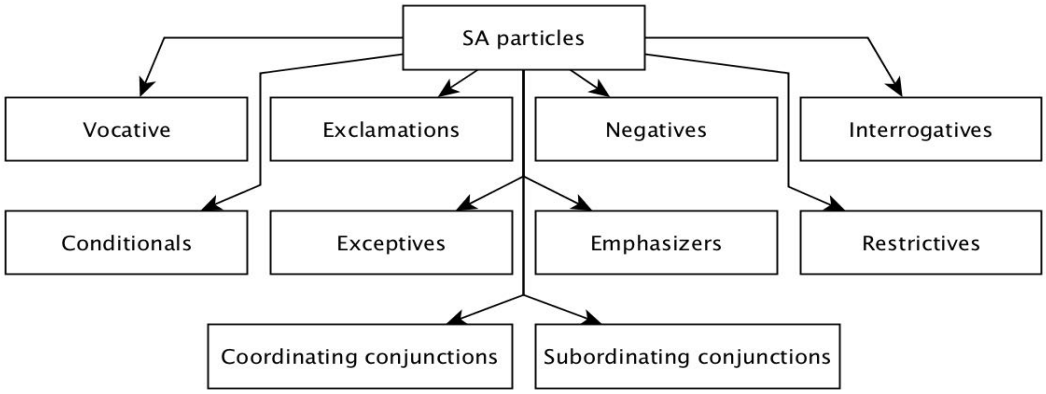


Figure 6.4: The main types of SA particles.

[58]. This term is the best one to describe the wide range of word classes that will be included in this experiment and is defined as a 'Wide-reaching term, including all indeclinable word classes such as adverbs, conjunctions, prepositions and other particle classes such as scalar particles, discourse markers, modal particles, negation, interjections'(Bussmann, 2006, p. 867)

[59] The distinction between prepositions and prepositionals was made by Badawi et al. (2013 p. 201). The former are entirely underived elements that have one function as particles of obliqueness, while the latter are nouns that have multiple functions as adverbials or are used as space and time qualifiers.

Furthermore, each type of particle is divided into several subcategories based on meaning or a particular function of specific particles. For instance, Table 6.21 shows several subcategories of exclamation particles in SA with examples.

Table 6.21: Examples of exclamation types in SA (Badawi et al., 2013, p. 44).

Type	Example		
bound exclamations	وازيداه	wazīdāh	Woe upon Zayd
free exclamations	كَلَّا	kallā	not at all
agreement or dissent	نعم	naʿam	yes
warnings	إياك	ʾiyyāk	be careful
surprise	هيهات	hayhāt	how remote
sorrow	وَيلٌ	waylun	Woe to
enthusiasm	هَلُمَّ	halumma	let's…
wishes	يا ليت	yā layt	would that, if only
command	هات	hāt	give it here
quantitative	رُبَّ	rubba	how much! how few!

However, this complex classification of particles emphasises the heterogeneous nature of linguistic behaviour in SA, and this complexity should be reflected in the extraction findings of prepositional AMWEs in this study.

After conducting the pre-processing task and preparing the data for the extraction experiment, the same experimental procedures were implemented as described in section 6.2. The linguistic phase in this extraction experiment involves several processing tasks, beginning with automatically annotating the corpus using the SAP toolkit. Then, based on frequency data and other resources as described in section 4.6.3, a list was generated of the most predictive morphosyntactic selection patterns of prepositional AMWEs. Table 6.22 shows several examples from the selection patterns extracted in this experiment; these show that the patterns include varieties of prepositional AMWE constructions which range from phrases with two tokens to six word expressions.

Table 6.22 Examples of notable patterns discovered for nominal AMWEs.

N-gram	Pattern	N-gram	Pattern
2	IN NN	4	IN NN NN DTNN
	IN DTNN		IN NN NN NN
	IN IN		IN NN DTNN DTJJ
	IN NNP		IN NN DTNN NN
	IN VBP		IN NN DTNN IN
	IN NNS		IN DTNN CC DTNN
	IN VBD		IN NN DTNN DTNN
	IN DTNNS		IN DTNN IN NN
	IN DTNNP		IN NN CC NN
	IN JJ		IN NN IN NN
	IN NN DTNN		IN NN DTNN CC DTNN
3	IN NN NN	5	IN NN NNP IN NN
	IN NN NNP		IN NN DTNN IN NN
	IN NN JJ		IN NN DTNN CC NN
	IN DTNN IN		IN NN NN NN DTNN
	IN DTNN NN		IN NN NN DTNN DTJJ
	IN NN VBP		IN NN NN DTNN NN
	IN DTNN DTJJ		IN NN NN NN NN
	IN NN IN		IN NN NN IN NN
	IN DTNN DTNN		IN WP VBP IN NN
6	IN NN NNP IN NN NNP		
	IN DTNN CC DTNN CC DTNN		
	IN NN DTNN PUNC CC VBD		
	IN NN DTNN CC NN DTNN		
	IN PRP CC NN CC VBD		
	IN NN NN DTNN CC DTNN		
	IN NN DTNN PUNC CC NN		
	IN NN DTNN PUNC CC VBP		
	IN NN DTNN IN NN DTNN		
	IN NN DTNN PUNC CC IN		

In the next stage, these patterns were used to extract several lists of more than 31.457k AMWE instances from the corpus which represent the multiple types of expression included in this study.

Regarding the treatment of discontinuity expressions, the same methods were implemented as described in previous studies which included using several search patterns and regular expression techniques to extract non-contiguous PAMWE candidates. Table 6.23 provides several examples of possible intervening words found in the extraction of the expressions بغض النظر عن As can be seen, the gap in this example ranges from a single token to a longer phrase with four tokens.

Table 6.23 Example of multiple intervening words in PAMWE candidates

last part	Intervening words		Initial part
عن	ايضا	aydan	بغض النظر
ʿan	عما تتمخض	ʿammā tatamaḫḫaḍ	biġaḍḍ annaḍar
	اخي العزيز	aḫī alʿazīz	
	عن تقارير تصدر	ʿan taqārīr taṣdur	
	عن جلب مشاهدين اكثر	ʿan jalb mušāhidīn akṯar	

The generated lists then underwent several filtering tasks using statistical and linguistic methods such as frequency threshold, word stop lists and NEs removal. Furthermore, for the bigram results in this study, the best AM scores were used based on the findings of the evaluation experiment reported in section 5.6. These were used to sort the extracted lists in ascending order based on AM scores, as shown in the examples in Table 6.24. This statistical data helped filter out many unimportant AMWE candidates with lower AM scores. However, for longer extracted phrases, counting of frequency data was used to exclude unwanted items from the initial AMWE lists that were extracted.

Table 6.24: Samples of bigram PAMWE instances sorted by MI and MI.log.F AMs[60].

AMWE bigram	MI score	AMWE bigram	MI.log.F score
بمعزل	4.412	بالإضافة	46.229
الوخز بـ	4.388	بالنسبة	43.423
إلى المرفقين	4.267	يتعلق ب	36.508
عوقب بـ	4.246	القيام ب	36.325

[60] [DTNN-DTJJ] and [NN-DTNN] morphosyntactic patterns used in the AMWE examples presented.

ناجمة عن	3.57	لكي	36.136
تلتحق بـ	3.568	بشكل	35.554
بحذافيرها	3.567	بسبب	35.004
بموجبها	3.567	بمعزل	32.072
زاخرة بـ	3.566	بواسطة	32.063
الاطاحة بـ	3.565	المتعلقة بـ	31.973

The various filtering tasks implemented in this experiment reduced the initially generated list to 15.678k candidates. Table 6.25 presents examples from the extracted list after the candidate filtering phase.

Table 6.25: Examples of used prepositional selection patterns with AMWE instances.

Pattern	MWE candidate	Pattern	MWE candidate
IN-IN	لكي	IN-NN-NN	ب شكل عام
	اذ لولا		ب صفة خاصة
	لأن		ب حد ذاته
IN-NN	بشكل	IN-NN-DTNN	ب غض النظر
	بسبب		في نفس الوقت
	بدون		ب عين الاعتبار
IN-DTNN	بالنسبة	IN NN NN	في تناغم مع الحالة
	بالإضافة	DTNN	ل دغدغة عواطف البسطاء
	بالتالي		ب بلوغ مراتب التميز
IN-VBP	فيما يتعلق	IN NN NN NN	ب دون قيود حديدية
	فيما يلي		ب عدد غير قليل
	فيما يبدو		ب طريقة غير مباشرة
IN-NNP-NNP	الى رضوان الله	IN NN DTNN	ب بالغ الحزن و الأسى
	على راسهم أمريكا	CC DTNN	ب قيم الترابط والاستقرار
	عن ابي هارون		ب عين الترفع والازدراء

As mentioned previously, several types of SA particles were included within the concept of prepositions in this extraction task. These variations can be seen in the generated list of instances. Table 6.26 shows several AMWE examples which represent expressions that start with multiple types of particles.

Table 6.26: Examples of AMWE starting with various types of particles.

Particles		Meaning	Candidate	Examples
اذا	ʾiḏā	if	اذا عرف السبب بطل العجب	ʾiḏā ʿurif assabab baṭal alʿajab
ما	mā	what	ما لا اذن سمعت	mā lā ʿuḏun samiʿat
متى	matā	when	متى استعبدتم الناس و قد ولدتهم امهاتهم أحرارا	matā istaʿbadtum annās wa qad waladathum ʿummahātuhum ʾaḥrār
أينما	ʾaynmā	wherever	أينما تولوا فثم وجه الله	ʾaynamā tawallū faṯamma wajhu ʾallāh
هنا	hanā	here	هنا وهناك	hunā wahunāk
إلا	ʾilā	except	إلا من رحم الله	ʾilā man raḥim ʾallāh

Finally, the extracted lists were classified into several homogeneous sets based on their linguistic characteristics such as the morphosyntactic structures and the number of tokens in the expression'. However, the results obtained in this experiment emphasise the importance of prepositional AMWEs and provide intensive corpus-based evidence for the common and most salient expressions of these types in SA. A sample from the final extracted lists from this study will be used in the following evaluation task. Table 6.27 provides a summary of the results of the three extraction experiments conducted in this study.

Table 6.27: Summary of the findings of the three experiments before and after the candidates were filtered.

AMWE type	Before the candidates were filtered	After the candidates were filtered	Total
nominal AMWE	13.287	14.572	52.243
verbal AMWE	24.267	13.287	37.554
prepositional AMWE	31.457	15.678	47.135
total	93.395	43.537	

6.6 Validation and evaluation

As described in section 2.2.4, several methods of evaluating MWE extraction in the literature have been suggested without any explicit agreement or preference for a specific method. This might be due to the various experimental contexts and the multiple existing interpretations of what is meant by valid MWEs in the literature. This heterogeneity leads to different views of the evaluation methodology; hence, claiming the value of any single standard of MWE evaluation is a somewhat controversial issue. Instead every researcher must select the most appropriate evaluation methods based on the context and the targeted applications of their work.

In this evaluation experiment, a quantitative evaluation was used based on the automatic and manual classification of a random sample of the extracted candidates. In these samples, a range of frequency levels, different MWE lengths, and multiple morphosyntactic constructs were included from the findings of the three extraction tasks reported in this chapter.

The selected evaluation method adopted in the study has been used in several similar MWE research studies (e.g., Da Silva et al., 1999; Seretan, 2011). However, it is important to mention that, in the current evaluation, use was made of the reference lists generated in the previous experiment reported in chapter 4 where 4651 validated AMWEs were extracted that underwent a manual validation task. Manual annotation was used in the evaluation due to the limited coverage of the available reference lists. Furthermore, in the manual annotation part of the classification task, the same procedures were followed as described in section 4.7.1, which includes detailed descriptions of the AMWE selection criteria and manual annotation validation testing and guidelines.

Based on the outputs of the classification task applied to the test datasets, the precision scores were calculated for each dataset along with the average precision for each extraction experiment. The reference data used in the classification evaluation of the extracted list were based on the previously manually evaluated list of true AMWEs occurring in SA corpora that was described in detail in chapter 4. Furthermore, following Krenn et al. (2004) and to eliminate the risk of subjectivity, it was important to ensure a certain degree of agreement among inter-annotators on the manual part of the classification task.

The evaluation samples of the extracted MWE candidates consisted of 15 test datasets which represent a variety of the AMWEs targeted in this study. These datasets were divided into five categories based on the number of grams in the candidates and into three classes based on whether the morphosyntactic patterns were nominal, verbal, or prepositional expressions. Random sampling from the final extracted lists of AMWEs was applied with a total of 7500 candidates distributed into 500 items in 15 classes. Table 6.28 presents a summary of information regarding the test datasets.

Table 6.28: Basic information about the evaluation samples.

Code	AMWE type		Code	AMWE type		Code	AMWE type	
NTS1	nominal	2-grams	VTS6	verbal	2-grams	PTS11	prepositional	2-grams
NTS2		3-grams	VTS7		3-grams	PTS12		3-grams
NTS3		4-grams	VTS8		4-grams	PTS13		4-grams
NTS4		5-grams	VTS9		5-grams	PTS14		5-grams
NTS5		6-grams	VTS10		6-grams	PTS15		6-grams

Thus, in the evaluation, the following steps were conducted:

- Generating an evaluation sample from the extracted candidates to reflect a range of frequency levels, pattern types, and phrase lengths. Table 6.29 presents examples of lexical items from the evaluation dataset samples. The datasets represent multiple morphosyntactic patterns and include lexical items ranging from 2-grams to 6-grams.

Table 6.29: Candidate examples from the evaluation datasets.

n-grams	Pattern	Candidate Examples	
2	IN-DTNN	عن النبي	ʿan annabī
		باللغة	billuġa
		بالعمل	bilʿamal
		بالتاكيد	bittaʾkīd
		بالذات	biḏḏāt
3	VBD-DTNN-DTNN	اشار الامين العام	ašār alʿamīn alʿāmm
		بدا العد التنازلي	badaʾ alʿadd attanāzulī
		اطلقت الرصاص الحي	ʾaṭlaqat arraṣāṣ alḥayy
		بلغت القلوب الحناجر	balaġat alqulūb alḥanājir
		القت الشرطة القبض	alqat aššurṭa alqabḍ

4	NN-NN-CC-NN	بين عشية و ضحاها	bayn ʿašiyya wa ḍuḥāhā
		تحية اجلال و اكبار	taḥiyyat ʾijlāl wa ʾikbār
		دون قيد او شرط	dūn qayd aw šarṭ
		امام مرأى و مسمع	ʾamām maraʾā wa masmaʿ
		مسالة حياة او موت	masʾālat ḥayāt aw mawt
5	NN-DTNN-CC-NN-DTNN"	تقرير المصير و اقامة الدولة	taqrīr almaṣīr wa iqāmat addawla
		رئيس الوزراء و وزير الدفاع	raʾīs alwuzarāʾ wa wazīr addifāʿ
		بداية المجتهد و نهاية المقتصد	bidāyat almujtahid wa nihāyat almuqtaṣid
		تحت الارض و يوم العرض	taḥt alʾarḍ wa yawm alʿarḍ
6	VBP-IN-NN-DTNN-CC-DTNN	يقودها ل بر الامان و الوحدة	yaqūdhā li barr alʾamān wa alwaḥda
		يؤدي الى اثارة الشك و القلق	yuʾdī alā aṯārat aššakk wa alqalaq
		نفضي الى رحلة التاكل و الضياع	nafḍī ʾilā riḥlat attākul wa aḍḍiyāʿ
		يؤشر الى نبض الناس و الشارع	yuʾaššir alā nabḍ annās wa aššāriʿ
		يعبرون عن مدى الحب و الولاء	yuʿabbirūn ʿan madā alḥubbi wa alwalāʾ

In the next step, the candidates in the sample were validated by automatic and manual annotation. In the automatic annotation, the extracted candidates were aligned to the gold-standard reference lists and the successfully matched items were automatically classified as valid AMWEs. Manual annotation was also used due to the lack of coverage in the reference list. In this part of the evaluation, two annotators completed the classification of phrases that were missed in the reference list. Several important issues were also addressed to ensure the quality of this task. For example, the coders were two experts in Arabic linguistics who had carried out research in this area and were provided with the adopted definition of AMWE accompanied by detailed annotation guidelines. Moreover, the degree of inter-coder agreement was tested using the kappa statistic κ (Cohen, 1960). More details about the evaluation and annotation, and the interpretation of the inter-annotator agreement test, is provided in section 4.7 of this thesis.

The kappa result in this experiment was 0.54, which is a moderate degree of agreement. However, in the literature, there is no consensus on a single cut-off or threshold point for measuring reliability using this test but, in general, the higher the score obtained, the more reliable the annotation task. The acceptance level for this test varies in the literature according to the purpose of each evaluation. However, with

vague and complex issues such as those in MWE classification, a low degree of agreement can be anticipated.

Based on the annotation findings, the precision scores for each test dataset were calculated and this is the primary evaluation figure in this study. Additionally, the mean average precision was calculated as follows to determine the overall performance of the AMWE extraction model.

$$Precision = \frac{\text{T annotated MWEs}}{\text{T annotated MWEs} + \text{F MWEs.}}$$

Tables 6.30 provides a summary of the annotation and evaluation results along with the precision scores for each dataset and the MAP measures for each type of evaluation data.

Table 6.30: Statistical information about the evaluation findings of the test datasets.

n-grams	**2**		**3**		**4**		**5**		**6**		**MAP**
NTD	428	0.86	467	0.93	431	0.86	324	0.65	277	0.55	0.77
VTD	429	0.86	409	0.82	303	0.61	328	0.66	177	0.35	0.66
PTD	325	0.65	378	0.76	317	0.63	320	0.64	118	0.24	0.58
MAP	1182	0.79	1254	0.84	1051	0.70	972	0.65	572	0.38	5031

The data shows the extraction model performs better with bigram and trigram candidates with a MAP of 0.79 and 0.84. In contrast, longer candidates of 5 and 6-grams obtain the lowest MAP scores with 0.65 and 0.38, respectively. The MAP scores based on the three types of AMWE show that the nominal test datasets achieved the best MAP score of 0.77, followed by verbal and prepositional test datasets with MAP scores of 0.66 and 0.58, respectively However, it was not possible to calculate the recall scores in the evaluation due to the limited coverage of the used reference data and the lack of information about all the true AMWEs in the used corpus.

6.7 Error analysis

The automatic and manual annotations conducted in the evaluation task revealed useful insights into the types of error found in the extraction outputs by analysing the false AMWE candidates. Table 6.31 provides a summary of the common errors found in the evaluation task with examples of the extracted instances from the test datasets.

Table 6.31: Examples of the main types of error in the evaluation datasets.

Class	**Error type**	**Example**	
tokenisation	split error	معدلات - ب - ـطالة - ف - ي	muʿaddlāt - bi - ṭāla - f - y
	miss-split	تتسلم{ها} - الوزارة	tatasllmu{hā} - alwizārat
POS tagging	noun	على- سبيل-المثال	ʿalā- sabīl-almiṯāl
	verb	ان ابواب	inn abwāb
	preposition	مع-هد- البحوث	maʿ-had- albuḥūṯ
semantic	meaningless expression	على- ك- ل	ʿalā- k- l
other	spelling	يصلي- ركة	yuṣṣalī- rakat
	word Order	الواقع –ارض- على	alwāqiʿ –arḍ- ʿalā
	ill formed structures	لعد – م -من	liʿad – m - min

A close analysis of these examples reveals the characteristics of the main types of error found in the evaluation datasets. Most errors stem from the automatic linguistic annotation implemented in the corpus by the SAP toolkit. For instance, in the first two error classes, several negative candidates were excluded due to tokenisation or POS annotation mistakes. In addition, the absence of short vowel representations in the written text leads in many cases to the wrong POS annotation, because the number of words in SA can be annotated with multiple POS tags based on their pronounced forms. For instance, the word كتب can be considered a noun or verb based on the type of short vowels used. Other kinds of errors stem from the semantics of the extracted expressions or spelling mistakes. Furthermore, several errors were found in the process of extracting instances for the selection of morphosyntactic patterns, such as the matching of ill-formed structures or selecting a construct in the wrong word order. Nevertheless, all these types of error provide informative insights about the

shortcomings of the AMWE extraction model which can be avoided in future work to eliminate the number of unwanted items in extraction outputs. Ultimately this improves the overall performance of the extraction model.

6.8 Summary of Results

The finding in this experiment for AMWEs extracted based on the most common syntactic structures demonstrates the complexity of this phenomenon in SA. However, in contrast to the idea that most MWEs are fixed phrases, the data obtained in the experiments reported in this chapter show the opposite as it was found that most AMWEs extracted represent a variety of morphosyntactic structures that undergo several types of modification at various levels of linguistic analysis.

In the first study on nominal AMWE, which is the most frequent type of MWE, the findings show multiple types of nominal constructs that cover multiple morphosyntactic patterns. Most valid candidates based on the evaluation datasets came in the form of 3-gram expressions which indicates the need for future work. The longer candidates that exceeded 4-grams were the lowest type of AMWEs; hence, this finding would be beneficial in analysing the linguistic behaviour of these lexical units and in the process of selecting predictive morphosyntactic patterns for nominal AMWE extraction tasks. In the second experiment on verbal AMWE, the findings showed that most extracted candidates were flexible types of structure that were affected by the various linguistic features. Thus, the extraction model for verbal AMWEs should permit a wide range of flexible selection patterns to improve the chance of discovering these types of expression. Finally, the various syntactic properties of the retrieved particle constructs in the study show the need for specific investigations of these lexical units in AMWEs; they should therefore be distinguished from other types of expression in the extraction process.

Furthermore, the overall findings of the experiments provided informative insights into the linguistic behaviour of AMWEs based on a large body of corpus-based evidence. The AMWEs found in most types of constructs in SA were nominal, verbal, adjectival, coordination, apposition preposition, apposition, and copular constructs. The semantic analysis of the extracted lists shows that they also belong to a variety of semantic fields including art science and education, and have a range of discursive

functions such as informational, modalising, and structural functions. Regarding semantic compositionality, although in this study there was an in-depth focus on this type of semantic analysis, multiple instances from the extracted data represent various levels of idiomaticity, starting from complete non-compositional candidates to others used in terms of their literal meaning. However, more corpus-based studies should be conducted to gain further insights into the linguistic and semantic behaviour of AMWEs.

6.9 Conclusion

In this research, multiple types of MWE have been extracted using a set of morphosyntactic selection patterns derived from various types of resources. Thus, three main AMWE extraction experiments were conducted based on a large annotated SA corpus and using a hybrid extraction model with several modifications within each experimental setting. The findings show that the use of linguistic and statistical components in the extraction task proved to be very useful in improving the overall discovery of multiple types of AMWs.

The results obtained in this chapter help to remedy the deficiencies in AMWE research by covering a wide range of AMWE constructions during the extraction process. This ultimately enhances the lexicon of AMWEs with new types of lexical units based on multiple morphosyntactic patterns. The evaluation used in this experiment shows that the shorter the target AMWEs, the better the performance achieved by the extraction model. These findings can be justified by the high frequency of those types of AMWEs in which the statistical methods generally work best. These results are also in line with those of other studies conducted on Arabic and other languages (e.g., Moirón, 2005; Attia et al., 2010; Bounhas and Slimani, 2009).

The finding reported in this chapter illustrate the need for further larger-scale AMWE extraction experiments to explore the characteristics of various MWEs in SA in depth. This is a particularly important task to be tackled for new language genres given the dominant use of user-generated content applications which provide new LRs to investigate various linguistic phenomena. Specifically, it will help in discovering new AMWE items used by the virtual interactive communities on social media.

7 A representational model for AMWE lexicon

7.1 Introduction

This chapter addresses the third research question of this thesis by describing the AMWEL computational representations at different linguistic levels based on international standards for representing various types of LRs.

Section 7.2 provides brief explanations about the core properties of the adopted representational model, and then section 7.3 lists the representations layers in the AMWE model based on multiple levels of linguistic analysis. The work presented in this chapter published in Alghamdi and Atwell, 2017.

7.2 Properties of MWE Computational representations

Based on the primary project objectives, the annotation scheme needed to be easy to integrate into different types of NLP systems, in line with state-of-the-art standards in lexical mark-up research. In addition, the adopted scheme could not be restricted to any particular grammatical framework as it needed to be reusable, as Odijk (2013, p. 189) emphasised:

> *'Lexical representations of MWEs that are highly specific to particular grammatical frameworks or concrete implementations are undesirable since it requires effort in making such representations for each new NLP system again and again and the degree of reusability is low'*

Another essential property of current representations is the flexibility which cuts across all types of AMWEs and covers discontinuous as well as contiguous phrases; it also needs to be human readable and equally adapted for NLP systems to accommodate different end users' needs. However, most of the previous studies on AMWE annotation schemes have prioritised certain types of expressions or language genres to the exclusion of others. Therefore, they are not appropriate for representing multiple kinds of AMWEs in the current lexicon which should allow for permutations

across various linguistic levels. The computational AMWE representations are encoded in Extensible Markup Language XML because this is the most flexible and the most used method in the formalism of computational LR. The final version will be converted into HTML pages so that the content can be published on the Internet.

This project also benefited from the international standard lexical markup framework (LMF) which was the result of the contributions of 60 experts who have worked for more than five years to develop lexical representations and standards for different types of computational LRs (Francopoulo, 2013; Francopoulo and Huang, 2014). The LMF describes the basic hierarchy of information of a lexical entry and also has specific provisions for MWEs, specifically a normative NLP MWE patterns extension, illustrated with examples in the form of a UML class diagram and XML hierarchy model (Francopoulo and George, 2008). It is important to note that adopting standardisation when building computational LR can be very beneficial, especially in NLP oriented applications. For instance, Francopoulo (2013, p. 3) states that:

> *'The significance of standardisation was thus recognised, in that it would open up the application field, allow an expansion of activities, sharing of expensive resources, reuse of components and rapid construction of integrated, robust, multilingual language processing environments for end-user '.*

Furthermore, the representations system developed pays particular attention to enriching the lexical entries with extensive linguistic information to allow for various types of end users and to prepare the LR for any potential use. Atwell (2008, p.4) states that 'For developers of general-purpose corpus resources, the aim may be to enrich the text with linguistic analyses to maximise the potential for corpus reuse in a wide range of applications.' In the following section, a brief description of the type of users targeted in the AMWEL project is presented. This is followed by a detailed illustration of the adopted AMWE classifications and representations across different linguistic levels.

7.3 AMWEL Computational Representations

As mentioned previously, in the design of lexicon annotation and classifications, this project takes into account the LMF core package and the extension of MWE patterns

with the necessary deviations to facilitate the reusability and connectivity of AMWEL to other LRs and various NLP systems and applications. This section describes the computational representations and the labels adopted for each class of MWEs and propriety property with examples from Arabic corpora.

As much use has been made of automated procedures as possible to reduce the time and effort involved in the annotation process. All the representations in the current version of this annotation scheme are classified into four main categories as follows: basic lexicon information, linguistic properties, pedagogical, and any other related information, which involves all the representations that do not belong to any of the previous three annotation groups.

7.3.1 Basic lexicon information

This class is mainly adopted from the MWE extension in the LMF framework and expresses the primary details on the AMWEL that can be useful for LR end users. The attributes in the global information class illustrate a brief abstract about the project which includes: label author, language coding, and script coding. Main Lexical Entry is the core class for each lexical entry and involves written form, related form, and lexicographic type. Other classes aim to represent the details of MWE components in their various linguistic manifestations.

Table 7.1: Basic lexicon information representations in AMWEL.

Class Name	Subclasses and attributes
Lexical Resource	
Global Information	Label
	Comment
	Author
	Language Coding
	Script Coding

As can be seen in Table 7.1, the ID attribute, which can be seen in most annotation classes, was created to facilitate the linkage between shared annotation classes; thus, it can be targeted by cross-reference links. The comments attribute is specified to provide any necessary information which might explain the annotation class. This information is encoded in XML; Figure 7.1 provides an example of the XML fragment of the Global Information class:

```
<GlobalInformation>
    <feat att="label" val="Arabic Multiword Expressions Lexicon"/>
    <feat att="comment" val="مدخل تفصيلي لخصائص التركيب في أمس الحاجة"/>
    <feat att="author" val="AymanAlghmdi"/>
    <feat att="languageCoding" val="ISO 639-3"/>
    <feat att="scriptCoding" val="ISO 15924"/>
</GlobalInformation>
```

Figure 7.1: An example of lexicon information annotated in XML.

7.3.2 Linguistic representations

The linguistic annotation classes are the core package of the AMWEL model and provide a detailed linguistic description of each ArMWE in the lexicon. The annotations are classified into six main layers; each one is dedicated to linguistic levels starting from the shallow orthographic form of the lexical entry to the in-depth semantic and pragmatic features of MWE. The following subsections present a brief explanation of these linguistic annotations.

7.3.2.1 Basic linguistic description

The first five classes provide the basic linguistic description of MWEs which was adopted from the MWE pattern extension model in LMF standards (Francopoulo, 2013), as shown in Table 7.2.

Table 7.2: Basic linguistic representations of MWE.

Class Name	Subclasses and attributes
Main Lexical Entry	Id
	Comment
	Written Form
	Related Form
	Lexicographic Type
List of Components	Component
	Related component

MWE Pattern	Id
	Written Template
	Comment
MWE Node	Syntactic Constituent
	Pattern Type
MWE Lex	Structure Head
	Rank
	Lexical Flexibility
	Graphical Separator

The Main Lexical Entry is the core class of each lexical entry and is associated with all the annotation features. It also has several attributes related to written and other forms of MWE. The lexicographic types of the expressions represented by several labels are presented in Table 7.3 with examples from the lexicon.

Table 7.3: Examples of lexicographic type labels in AMWEL.

Lexical Types labels	Examples	Translation
Compound noun	عيادة الطبيب	Medical Practice
Support verb	طفح الكيل	Fed up
Quotation	ضرب صفحاً	Ignore
Idiom	مقطوع من شجرة	Cut from a tree
Proverb	ضرب الحديد وهو حامٍ	Hit the iron while it is hot

The MWEs pattern is a shared resource which provides information about different lexical combination phenomena. This class is associated and explained by the list of components that contain all the constituent expression words. The node classes represent the structural properties of the given phrase by providing information on syntactic constituent and pattern type. The first feature illustrates the written template form of the structure, for instance, the syntactic components of the English phrase to take off is Verb_ Preposition or VP; an equivalent Arabic example can be seen in the phrase, *ʾaḫaḏ ʿan* أخذ عن which is also classified as VP structure. In Table 7.4, examples of syntactic constituents found in AMWEL are listed.

The pattern type represents the degree of morphological, lexical and grammatical flexibility of phrases by using a scale of three levels, as illustrated in Table 7.5.

Table 7.4: Examples of the classification of syntactic constituents in AMWEL

Label	Example
Noun_Noun	تكميم الأفواه, takmīm al'afwāh
Verb_Noun_Preposition_Noun	تجمد الدم في عروقه, tajmad addam fī ʿurūqih
Noun_Adjective	اليد المغلولة, alyad almaġlūla
Noun_Adverb	الأيام بيننا, al'ayyām baynanā
Noun_Preposition	التغطية على, attaġṭiya ʿalā
Preposition_Noun_Preposition	من أجل أن, min ʾajl ʾan
Noun_Preposition_Noun	النوم في العسل, annawm fī alʿasl

Table 7.5: Classification of pattern types with Arabic examples.

Flexibility degree	Example
Fixed MWE	رجع بخفي حنين, rajaʿ biḫuffay ḥunayn
Semi-fixed MWE	أثلج/أثلجت صدره/صدرها, ʾaṯlaj/ʾaṯlajat ṣadruh/ṣadrahā
Flexible MWE	أثقلته/أثقله/أنهكته الأعباء/الحمل/المسؤوليات، ʾaṯqalath/aʾṯqalah/ʾanhakath alʾaʿbāʾ/alḥiml/almasʾūliyyāt

The MWE 'lex' class is used to provide a reference to each lexical component in the list of components. It also provides lexical classifications of each list of components based on the possibility of allowing some substitutions in the lexical items. Hence, two values are specified for each component: one for MWEs that can be alternated with other lexical items and the second for other MWEs that have to be used with the same lexical items or what are termed fixed MWEs. The Structure Head represents the first POS tag for the phrases, and the rank attribute shows the components order and any potential alternative orders. This feature is mainly essential for Arabic which has a high degree of flexibility in the word order within sentences. For instance, the MWE أقبلت عليه الدنيا *ʾaqbalat ʿalayh addunyā* has six possibilities for component order, as shown in Table 7.6.

Table 7.6: An example showing the flexibility of component order in AMWEs.

A	الدنيا	2	عليه	3	أقبلت	1
B	عليه	3	الدنيا	2	أقبلت	1
C	أقبلت	1	الدنيا	2	عليه	3
D	الدنيا	2	أقبلت	1	عليه	3
E	أقبلت	1	عليه	3	الدنيا	2
F	عليه	3	أقبلت	1	الدنيا	2

7.3.2.2 Orthographic representations

As described in Table 7.7, the orthographic annotation contains five attributes which in turn have several values. Three attributes express the orthographic variety of the expression, which can be very useful, particularly for NLP oriented users as it enables them to extract the LR in various formats according to the targeted NLP or ML tasks. An example of these types of representation can be seen in the phrase أعياه الأمر *ʾaʿyāhu alʾamru* which can be represented in various forms based on its orthographic features, as shown in Table 7.8.

Table 7.7: The linguistic annotation layers of AMWEL.

Class Name	**Subclasses and attributes**
Orthographic Features	Id
	Comment
	DIN31635RenderingInPlainEnglish
	Normalised Form
	Different Spelling Form
Phonological Features	Id
	Comment
	Diacritisation
	Phonetic Form
	Phonological Variants

Morphosyntactic Features	Id
	Comment
	Word Form
	Root
	Derivation form (Lemma)
	Stem
	Morphological scheme
	Part of Speech
	Grammatical Features
	Syntactic function
Semantic Features	Id
	Comment
	Sense
	Semantic Fields
	Idiomaticity Degree
	Semantic Relations
Pragmatic Features	Id
	Comment
	Usage Type
	User Type

Table 7.8: An example of the orthographic features of MWE أعياه الأمر , *exhaust.*

Orthographic Features	Expression example
DIN31635RenderingInPlainEnglish	ʾaʿyāh alʾamr
Normalised Form	اعياه الامر
Different Spelling Form	أعياه الأمر

For an example of the previous annotation in XML, Appendix H illustrates the XML fragment which represents the ArMWE فِي أَمَسِّ الْحَاجَةِ, *fī ʾamassi alḥājati, in urgent need.*

7.3.2.3 Phonological representations

At the phonological layer of annotation, a complete diacritisation of each phrase is provided which is an essential feature used in Arabic phonology to express the most common pronunciation form of AMWEs in SA. This representation is also particularly important because of the absence of short vowel symbols in Arabic script, which also plays a prime role at the syntactic and semantic analysis levels of the

lexical units. Other attributes are devoted to representing other phonological variants when available and a representation of the expression in IPA phonetic script.

7.3.2.4 Morphosyntactic representations

For the morphosyntactic representations, a modified version of LMF morphological patterns extension was used to provide detailed descriptions of the morphosyntactic feature of the phrase. This level of annotation is essential, particularly for Arabic which has powerful derivational morphological features that result in different variations for each word that will be represented in the AMWEL lexicon. Regarding the POS feature, components of expressions are classified into five categories according to their POS tag. Table 7.9 shows the adopted morphological tag set with MWE examples of the headword POS.

Table 7.9: Examples of the POS tags used in the morphosyntactic representations.

POS tag	Example
Noun	البرج العاجي alburj al'ājī
Verb	التزم الصمت iltazam aṣṣamt
Adjective	جنون العظمة junūn al'aḏama
Adverb	بين الحياة والموت bayn alḥayāt walmawt
Preposition	على قدم المساواة 'alā qadam almusāwā
Interjection	يا غالب يا مغلوب yā ġālib yā maġlūb

The morphological features for each component are represented in a specific element. However, the morphological properties are essential and useful information to include in the representations of MWEs because of the derivational and inflectional nature of Arabic morphology which means that words in Arabic are derived from specific roots; usually inflected words that share the same root belong to a common semantic field. This feature therefore helps to classify with ease all the words belonging to the same root into semantically similar groups based on the common morphological root. Table 7.10 shows an example of an Arabic root with its morphological patterns and inflection forms.

Table 7.10: Examples of morphological patterns and meanings of the root (s—m-ʿ).

Morphological patterns		Meaning
سمع	samiʿ	Listen (Past tense verb)
يسمع	yasmaʿ	Listens (Present tense verb)
اسمع	ʾismaʿ	Listen (Imperative verb)
مسموع	masmūʿ	Heard
سماعة	sammāʿa	Speaker (for computer, etc.)
سامع	sāmiʿ	Listener (Singular)
سامعون	sāmiʿūn	Listeners (Plural for male)
سامعات	sāmiʿāt	Listeners (Plural for female)

The grammatical features class represents four main properties: number, gender, tense for verbs, and person. Consequently, all these features involve several values which are described in detail in the grammatical properties of each MWE component. Table 7.11 provides examples of these linguistic features in Arabic.

Table 7.11: Examples of the annotation of grammatical features.

Grammatical features	Values
Number	Signal, plural
Gender	Male, female, things
Tense	Past, present, imperative
Person	Third person

7.3.2.5 Semantic representations

This level of annotation constitutes four main classes created to represent the semantic information of MWEs. The ‘Sense Set' class represents the variations of meaning of MWEs in different contexts that are associated with a corpus example that reflects the real use of the phrase. The ‘Semantic Fields' class groups the phrases into several categories based on the main semantic fields. The semantic tagset developed at Lancaster used in representing various types of AMWEs, the tagset consists of 232 semantic tags based on 21 main classes in the adopted taxonomy as described in section 2.4.9 of this thesis.

The idiomaticity degree feature classifies the MWEs into three categories based on the ambiguity levels of the phrase as follows: full opaque, semi-opaque, and compositional MWEs. Fully opaque MWEs involve expressions where there is no

semantic relation between the general meaning of the phrase as a whole and its component parts, such as:

عَلَى كَفِّ عِفْرِيتٍ

ʿalā kaffi ʿifrītin

عَلَى قَدَمٍ وَسَاقٍ

ʿalā qadamin wasāqin

طَالَتْ أَظَافِرُهُ

ṭālat ʾaḏāfiruhu

Semantic Relations is a class representing the oriented relationship between Synset instances, where three types of relations are included: synonymy, antonymy, and polysemy.

7.3.2.6 Pragmatic representations

The pragmatic annotation of MWE adds usage labels to MWEs that demonstrate the type of potential users or the possible situations in which this phrase can be used, such as academic, formal, and informal uses of the MWE. These features help in the deep understanding of an MWEs' pragmatic behaviour.

7.3.3 Pedagogical representations and other features

These representations aim to make the most of AMWEL in any language pedagogy related applications. Thus, this class provides valuable information that includes frequency attributes which show the degree of popularity of the phrase. In addition, the source label presents information about the source LRs where phrases were extracted.

The date label indicates the date of compiling the source corpus while the style label refers to the type of language genre such as standard, classical, or other Arabic dialects. The type element represents whether the MWE was from a written or speech corpus.

As listed in Table 7.12, the final class of the representations model was created to include all the information beneficial for LR end-users that cannot belong to any of

the previously described annotation classes. For instance, the status of annotation compilation for each lexical entry and the MWE equivalent in Arabic dialects or the translation of MWE in other languages.

Table 7.12: Pedagogical representations and other features of MWEs.

Pedagogical Features	Id
	Comment
	Learnability Levels
	Frequency
	Language Type
	Voiced example
	Language Source Name
	Language Source Link
Other Features	Id
	Comment
	Translation Equivalent
	Dialectic Equivalent
	Entry Status Levels

7.4 Summary

This chapter presented a detailed description of the lexical representations model that was applied in the development of a comprehensive AMWE lexicon for NLP. The model built upon previous attempts and standards in the computational lexical representations of MWEs; moreover, several innovative annotation features were added that enhance the usefulness and usability of AMWEL in various practical applications in NLP and LP. This work is a crucial and essential step towards more advanced and comprehensive research on the computational treatment of AMWEs.

8 Conclusions and Future Directions

8.1 Thesis summary

At the end of this journey, which has explored AMWEs from various perspectives, this chapter will end the thesis by presenting a summary of the literature and highlighting the main contributions of this project. It will also discuss the limitations of the research along with potential applications and future work. However, as illustrated in the introduction to this thesis, it is important to reemphasise -based on the findings of the multiple experiments conducted in this work- that AMWEs are complex and heterogeneous linguistic phenomenon which poses various problems for NLP computational tasks. These problems are more challenging in SA because of its distinctive linguistic features and the rich morphological system. So far in this thesis, a step has been taken towards improving AMWE computational tasks by implementing several AMWE extraction models to create an intensive AMWE lexicon that can be used in several NLP tasks. The LR developed in this thesis should pave the way for many subsequent projects that aim to enhance the computational treatment of this linguistic phenomena. However, through the many research phases in the project, it has become clear that, the deeper one delves into this phenomenon, the more complex and heterogeneous the nature of AMWE appears to be.

The research journey in this area is far from complete; much more work is needed in this area to address the many open research problems in MWEs and AMWEs in particular.

8.2 Literature summary

Following the introduction, in chapter three the conceptual framework for AMWEs was described by elucidating crucial theoretical issues, starting with providing a general background on SA and the motivation for selecting this specific variant of Arabic as the subject of this research. Furthermore, a brief linguistic description of SA was presented at various linguistic levels. The core concepts used in the study were then illustrated with a focus on the AMWE concept given the specific scope and

context of the research. The chapter also presented a brief description of AMWE characteristics and variants at various linguistic levels as well as surveying the existing typologies and classifications of MWEs with particular emphasis the adopted typology of AMWEs.

In the literature review chapter, a set of related works within various areas of research under four main topics was surveyed: AMWE discovery methods, MWE LRs, computational representations, and applications. In the first part, related work was discussed on extracting multiple types of MWE from corpora found in the literature. The research in this area was grouped into three main paradigms based on the kind of methods used in MWE discovery process. In addition, there was a brief review of existing evaluation methods used in various MWE extraction models. In the second part of the review, existing MWE LRs were discussed with a focus on AMWE LRs. In the third part, a survey of related work on establishing computational representations and the annotation of multiple types of MWEs was presented. However, it is important to note that in all the previous research areas covered in the literature, the focus was on the most relevant and important research related to this thesis. Furthermore, priority was given to reviewing and discussing related AMWE research in all the previous areas when it became available.

8.3 Research questions and objective revisited

At the beginning of this thesis, three central questions related to MWEs in SA were posed, which this project aimed to address. These were:

RQ1: From the perspective of NLP applications, which type of MWEs should be given priority?

RQ2: How can lexical units of the type defined in RQ1 be discovered by a computational extraction model?

RQ3: What are the standards and methods of best practice for linguistic annotations and computational representations of AMWEs at various linguistic levels?

Within these questions, the project also set out several research objectives which were as follows:

To develop a computational corpus-informed AMWE lexicon that can be incorporated into various Arabic NLP applications.

To establish standards for describing and encoding lexical entries in AMWEs at different linguistic levels (morphological, syntactic, lexical, and semantic).

To determine the information and annotation that will best serve the needs of language-related applications.

To propose an overall model for AMWE identification and extraction that will best suit the primary objectives of this research.

To explore the feasibility of creating an intensive AMWE LR by conducting several AMWE extraction experiments and constructing an intensive lexicon consisting of various types of AMWE entries with rich linguistic annotations.

These research objectives and questions formed the basis of various in-depth theoretical and experimental research studies on AMWEs that were reported in five chapters of this thesis. A summary of the main contributions and the efforts made to answer the research questions are provided in the following subsections.

8.3.1 Thesis contributions

In the following subsections, the conclusions drawn from our various research studies conducted in this thesis are summarised. These are divided into four main areas according to the research questions.

8.3.1.1 The theoretical framework for AMWE

In chapter two of this thesis, a detailed conceptual framework for AMWEs and their variation potential at multiple linguistic levels of analysis was presented. A review of several existing typologies of MWEs was undertaken and the most distinctive linguistic properties of AMWEs were elucidated. Based on corpus-based evidence and the results of empirical work conducted in related research areas, the general framework of AMWEs described in this thesis paved the way for the research tasks undertaken in the project by establishing the boundaries, context, and scope of the adopted conceptual framework of AMWE. The framework described in the thesis was not based on any pre-existing AMWE LRs nor was it related to any specific linguistic theories or computational formalisms. Instead, it was sufficiently general to cover a broad range of morphosyntactic constructs in SA. This framework can be used

in future work and will be beneficial for any related research on various aspects of opening problems within AMWE research areas.

8.3.1.2 AMWE extraction models

One of the most challenging tasks in the computational treatment of MWE is the automatic discovery and identification of MWEs in running text; in this thesis, the first task addressed was related to the extraction of multiple types of AMWEs from a large SA corpus to answer RQ2. Hence, several extraction experiments were implemented based on hybrid computational models that integrated statistical and linguistic techniques in the discovery process of AMWEs. The related work was reported in chapters four, five, and six. In chapter four, a hybrid model was used to extract initial reference lists with a broad coverage of AMWE variations that were then used as golden standard lists in the subsequent extraction experiments. Chapter five presented several empirical experiments that evaluated a set of AMs used in the extraction model of bigram AMWE candidates. The aim was to enhance the AMWE extraction model by using the best AMs to predict true AMWE items. Chapter six extended the extracted lists of AMWEs by taking advantage of all the previously conducted extraction experiments to explore the feasibility of using a wide range of morphosyntactic patterns in the AMWE extraction models. The AMWE extraction models implemented in this study along with the evaluation findings for each experiment provide valuable contributions to the fields of AMWE and, more generally, ANLP. For instance, the extraction models and the evaluation procedures can be replicated and used to extract AMWEs in various contexts and can also be applied to varieties of Arabic language text genres that were not covered in the thesis.

8.3.1.3 AMWE lexicon

The primary objective of the thesis was to build a large intensive AMWE LR that can be used to improve various NLP tasks. In this thesis an LR was developed that contained more than 10k AMWEs that were not restricted to any morphosyntactic constructions or semantic fields and were manually evaluated. This LR assisted with a comprehensive computational representational model that could enhance the usability and scalability of the lexicon developed for AMWEs. The lexicon developed is a valuable LR which meets the demands of related research on AMWEs, especially for evaluation studies, as Farahmand et al. (2015, p. 29) point out that ‘scarcity of

multiword expression datasets raises a fundamental challenge to evaluating the systems that deal with these linguistic structures'. At the time of writing this conclusion, the lexicon developed in this thesis is continually being enhanced and improved by adding more lexical items and enriching its linguistic annotation.

8.3.1.4 A representational model for AMWE knowledge

Chapter seven presented in detail the representational model of AMWEs at various linguistic levels, which is a necessary step in representing the linguistic analysis of AMWE knowledge in computational formalism. In the design of the lexicon model all the previous efforts in representing MWE LRs were taken into consideration with a particular focus on describing the distinctive linguistic properties of AMWEs. The representational model was designed to include a wide range of AMWEs and to be open to numerous extensions and improvements in future work to cover a complete linguistic description of the various AMWE types included in the lexicon. Furthermore, to ensure reusability the representational model does not adopt a specific linguistic or grammatical framework. Thus, the representational system developed can be reused and applied in various contexts and NLP tasks. This model provides a new contribution to related research areas because it presents a comprehensive formalism for representing a broad range of AMWEs and related linguistic knowledge.

8.4 Potential applications for AMWE LR.

The lexicon of AMWEs that was developed can be used and evaluated in various types of NLP applications. Furthermore, the LR developed in this thesis can have beneficial implications in other language-related domains such as linguistics, translation, lexicography, and LP research. In the following subsections, several examples are provided in which the availability of MWE LR plays a primary role in improving output quality and increases linguistic precision in the computational treatment tasks of natural languages. These applications present intriguing ideas and broad-coverage research opportunities in the field of AMWEs.

8.4.1 NLP related applications

The availability of AMWE LR is an essential step towards achieving a high quality precision output in most NLP tasks. Hence, in the following subsections, several examples of the potential applications of the developed AMWE lexicon will be briefly highlighted as worthy of consideration in future work. However, the focus is only on applications that have been applied to SA or other languages and have obtained significant findings.

8.4.1.1 Machine translation

MT is one of the most interesting and active research areas of NLP. Although several advances have recently been made, the translation of MWEs still faces several challenges in this area, especially when translating from and to morphologically rich languages. Thus, the use of MWE LR has proved to be beneficial in improving the overall performance of MT systems. Various methods have been suggested in the literature for integrating MWEs into MT. For instance, Pal et al. (2010) merges MWE knowledge into a Moses English–Bengali system as a pre-processing task. It does this by considering several types of MWE constructs as a single token in the implemented tokenisation scheme. Consequently, the systems 'translation' output has exhibited significant improvements in the quality and accuracy of the text being processed. In another study, Ren et al. (2009) integrated MWE into a phrase-based MT system by automatically extracting bilingual MWE and using an additional feature to represent phrases considered to be MWEs; they reported an encouraging and motivated improvement in MT performance by applying this strategy of MWE integration. In AMWE research, Carpuat and Diab (2010a) implemented what they called static and dynamic integration methods of AMWE into a statistical MT system to evaluate the usefulness of AMWE LR. The findings of their experiments also show an overall improvement in Arabic-English MT.

8.4.1.2 Language parsing

MWE knowledge is a fundamental part of natural languages, as described in detail in section 1.2.1. The inclusion of this knowledge is essential and positively influences overall parsing accuracy. This is evidenced by the findings of many research studies in the literature. For instance, Korkontzelos and Manandhar (2010) integrated MWE

knowledge into a shallow parsing task and found an increase of between 7.5% and 9.5% in the accuracy of processing text with MWEs. Wehrli et al. (2010) also embedded MWE knowledge into a language parsing task and found this resulted in substantial improvements over the standard method. The MWE LRs can be integrated into and have a positive impact on most language parsing levels, from the tokenisation task to the deep linguistic processing and morphosyntactic analysis. Constant et al. (2017, p. 862) state that MWE-aware parsing has three main benefits, which are '(1) to improve the syntactic parsing performances on sentences containing MWEs (both on internal MWE structure and on the surrounding sentence structure), (2) to improve MWE identification performance, and (3) to improve MWE discovery performance'.

8.4.2 Other applications

The AMWE LR developed in this thesis will also be of interest to researchers in other language-related areas, especially in LP, first and second language acquisition, and applied and theoretical linguistics research. An enormous amount of research has been published on the inclusion of MWE knowledge in these research areas, as will be briefly mentioned in the following subsections, which show the importance of MWEs in two examples: LP and linguistic applications.

8.4.2.1 LP applications

In language education research, particularly in the area of first and second language acquisition, MWE received early attention from researchers because of the significant effect MWE knowledge has on these areas. For instance, studies in first language acquisition found that children start learning languages by acquiring a mass of formulaic phrases they can reuse to express various meanings (Clark, 2008; Bannard and Lieven, 2012).

In second language learning, research asserts the positive influence of including MWE and FSs knowledge in the improvement of second language learning, especially for advanced learners. Hence, the availability of AMWE LRs can be of significant benefit in raising awareness of this linguistic phenomenon. Particular types of MWE, such as prepositional MWEs, have been found to be challenging to learn, especially for non-native speakers as reported in a series of error analysis studies of second language learners (e.g., Leacock et al., 2014; Leacock et al., 2010). Such learning requires

explicit LRs which illustrate these types of MWE in various language contexts. However, MWE knowledge can be integrated into and was found to have a positive impact on most stages and for any language learners in the learning and teaching process. For this purpose, several MWE LRs have been developed to enhance the presence of MWE in LP applications (e.g., Martinez, 2011; Durrant, 2008).

8.4.2.2 Linguistic applications

Access to an extensive lexicon of MWEs which represents their varying potential has interested linguists in exploring this phenomenon from different perspectives. This in turn can inform a better understanding of MWEs and their various manifestations at various linguistic levels. Descriptive and corpus-based studies benefit from MWE LRs because they provide them with intensive datasets that can be analysed and explored in the context of various research problems.

8.5 Study limitations

As has been mentioned throughout this thesis, in the MWE research area there are still many problems that have yet to be resolved due to the complex, heterogeneous and idiosyncratic nature of this phenomenon in all morphologically rich languages, such as SA. Although in this research strenuous efforts were made to overcome several limitations, as in every research project there were still several shortcomings that will now be discussed briefly.

8.5.1 Data sources

Although a large SA corpus was selected as the primary source of data for the experiments conducted in this thesis, developing a more carefully compiled, special, and representative corpus is an option that might lead to better findings. However, the use of a web-based corpus is not without limitations, such as the over or under representation of several types of language which will negatively influence corpus compilation procedures. However, because of time constraints, this research relied on a previously developed corpus because developing a new standard and reference corpus usually requires substantial financial resources and time and may quickly become out of date due to rapid changes in languages.

8.5.2 Linguistic analysis and annotation

Several limitations in the research stem from the use of an automatic SA linguistic toolkit in the experiments (e.g., SAP and MA) for morphosyntactic annotation and disambiguation analysis. Their final output and analysis has a limited degree of accuracy due to the nature of SA and the limited capacities of these computational tools. An ideal solution might be to carry out manual annotation, conduct a manual evaluation of the results, or build a better toolkit for SA analysis. However, these were beyond the scope of our study due to related constraints and the large size of the corpus used in the study.

8.5.3 Experimental setting and scale

Every experiment conducted in this thesis could be implemented in a different and perhaps more appropriate setting and could also be scaled to a broader context that ultimately supports the researcher's claims and generalisations of various aspects of AMWE phenomena. However, like the other limitations explained in this study, several restrictions had to be imposed for practical reasons. In summary, although the work reported in this thesis provides several valuable contributions to the relevant research fields as illustrated in section 8.3.1, it is fair to assert that in every research task implemented in this project, there is still substantial scope for various possible improvements and extensions. Remaining up-to-date is a challenging task for researchers, especially in computational and linguistic areas where rapid advances and an increase in the number of tools can become available within the blink of an eye in an era of big data and information explosion.

8.6 Future directions and open research problems

More theoretical and applied research is needed on AMWEs. Therefore, based on the main issues discussed in the study, several ideas for possible future work and extensions of the research will now be presented.

8.6.1 A theoretical framework for AMWEs.

MWE theory in SA is still in crucial need of further linguistic studies to explore its various theoretical aspects based on corpora and other LRs. Research can inform a comprehensive understanding of the multiple and varying potential of AMWEs and

the linguistic behaviour of this phenomenon in SA. Such theoretical corpus-based studies are an essential step in laying a solid theoretical foundation for advancing the computational treatment of MWEs at various linguistic levels. For example, future work might focus on comparative research between MWEs in SA and Classical Arabic or between SA and different Arabic dialects. Other research might compare the behaviour of MWEs in Arabic and other modern languages such as Spanish, French, or English.

8.6.2 AMWE computational tasks

In this thesis, the use of multiple hybrid AMWE discovery models to extract various types of AMWEs was investigated. However, in comparison to research conducted on English MWE extraction, the research problems related to AMWE in this area are far from resolved. Hence, much more research is needed on two main AMWE computational tasks: the discovery of new candidates and identification of AMWE items in running text. The following is a list of potential work in this area:

Investigating the use of various ML and DL techniques in discovering and identifiying AMWEs.

Exploring the use of semantic similarity methods based on contextual information in discovering new AMWEs and measuring their degree of compositionality.

Proposing new valid methods for evaluating large-scale discovery models instead of the current dependence on standard golden LRs or the selection of a sample dataset for the evaluation task.

Developing broad coverage AMWE identification models based on various supervised and unsupervised methods.

8.6.3 Extending the scale of the lexicon and enhancing it with rich linguistic annotations

The time constraints in the project prevented the size of the lexicon from being extended and from completing the computational representations and linguistic annotation of all the lexical entries in the LR developed. A short-term future task would therefore be to extend the current LR with new items and complete the rich annotation of the lexicon. Moreover, extensions and continuous improvements to the

representational model of the lexicon are essential to enhance its usability and scalability.

8.6.4 Integrating AMWE into NLP and LP applications

The ultimate goal of building an AMWE LR is to improve the computational treatment of this linguistic phenomenon; thus, future research should focus on integrating the AMWE lexicon into various NLP and language-related applications to improve their final output and to evaluate the developed LR. A wealth of research has shown that integrating MWE knowledge into NLP applications, especially in language parsing and MT, has a positive impact on reported performance, as described in section 3.4.

8.7 Summary

In this thesis, three research problems were investigated that were related to the complicated phenomenon of AMWEs in a specific time frame with limited access to LRs and computational tools and several other constraints that usually accompany similar PhD projects. Efforts were made to overcome many of these obstacles to find the best possible methods for deriving comprehensive answers to the research questions, as reported in this thesis. Nevertheless, at no stage of the research can it be claimed that the most valuable and final answer was obtained. However, the research capabilities and efforts made throughout this project meant that the primary specified objectives of this thesis were achieved. Research on MWEs is increasingly becoming multidisciplinary in nature which means researchers will benefit significantly from the work of interdisciplinary research teams throughout the world. This is particularly important when considering the unprecedented availability of linguistic data available on user-generated content platforms, such as social media apps (e.g., Twitter, Facebook, Instagram). Research on MWEs and in any language-related areas will find invaluable new sources of language data that will eventually open the doors to a wealth of research ideas that will subsequently result in considerable improvements in NLP computational processing and applications. In this respect, the project reported in this thesis forms a small part of larger-scale contributions towards achieving the long-standing dream of humanising and naturalising the use of machines.

REFERENCES

Abandah, G.A., Graves, A., Al-Shagoor, B., Arabiyat, A., Jamour, F. and Al-Taee, M. 2015. Automatic diacritization of Arabic text using recurrent neural networks. International Journal on Document Analysis and Recognition (IJDAR). 18(2),pp.183–197.

Abdou, A. 2011. Arabic Idioms: a corpus-based study. London: Routledge.

Al-Haj, H., Itai, A. and Wintner, S. 2014. Lexical Representation of Multiword Expressions in Morphologically-complex Languages. International Journal of Lexicography. 27(2),pp.130–170.

Al-Sabbagh, R., Girju, R. and Diesner, J. 2014. Unsupervised Construction of a Lexicon and a Repository of Variation Patterns for Arabic Modal Multiword Expressions In: Proceedings of the 10th Workshop on Multiword Expressions (MWE). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 114–123.

Alghamdi, A. and Atwell, E. 2018a. An Arabic corpus-informed list of MWEs for language pedagogy In: O. L. Dong, J. Lin, W. Xiao, M. Geraldine and P.-P. Pascual, eds. TALC 2018 13th Teaching and Language Corpora Conference. Cambridge, pp. 38–41.

Alghamdi, A. and Atwell, E. 2016a. An empirical study of Arabic formulaic sequence extraction methods In: The10th International Conference on Language Resources and Evaluation. Portorož, Slovenia: LREC.

Alghamdi, A. and Atwell, E. 2016c. Computational Identification and linguistic Classifications of Arabic Formulaic Sequences In: The 9th Saudi Students Conference. University of Birmingham, UK.

Alghamdi, A. and Atwell, E. 2017. Towards Comprehensive Computational Representations of Arabic Multiword Expressions In: R. Mitkov, ed. Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, pp.415–431.

Alghamdi, A. and Atwell, E. 2018b. Constructing a corpus-informed Listing of Arabic formulaic sequences ArFSs for language pedagogy and technology. Accepted paper submitted to International Journal of Corpus Linguistics.

Alghamdi, A. and Atwell, E. 2016b. Towards a Computational Lexicon for Arabic Formulaic Sequences In: The International Conference on Information and Communication Technologies. Rabat, Morocco: Information Systems and Communications Centre (CEISIC).

Alghamdi, A.A. 2015. The development of an Arabic corpus-informed list of formulaic sequences for language pedagogy In: The eighth international Corpus Linguistics conference. Lancaster: UCREL, Lancaster University, pp. 362–364.

Alnajadat, B.M. 2017. Pro-drop in Standard Arabic. International Journal of English Linguistics. 7(1), pp.163–172.

Alrabiah, M., Alhelewh, N., Al-Salman, A. and Atwell, E.S. 2014. An empirical study on the Holy Quran based on a large classical Arabic corpus. International Journal of Computational Linguistics (IJCL). 5(1),pp.1–13.

Altamimi, M.I. 2015. Arabic Pro-Drop. Master's Thesis, Eastern Michigan University.

Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A. and Suchomel, V. 2014. arTenTen: Arabic corpus and word sketches. Journal of King Saud University-Computer and Information Sciences. 26(4),pp.357–371.

Artstein, R. 2017. Inter-annotator Agreement In: Handbook of Linguistic Annotation. Dordrecht: Springer Netherlands, pp. 297–313.

Artstein, R. and Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics. 34(4),pp.555–596.

Attia, M. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks In: Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK.

Attia, M., Kayadelen, T., Mcdonald, R. and Petrov, S. 2017. PoS , Morphology and Dependencies Annotation Guidelines for Arabic. Google Inc California, USA.

Attia, M., Pecina, P., Toral, A., Tounsi, L. and van Genabith, J. 2011. An open-source finite state morphological transducer for modern standard Arabic In: Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing. Springer, pp. 125–33.

Attia, M., Toral, A., Tounsi, L., Monachini, M. and Genabith, J. Van 2005. An automatically built Named Entity lexicon for Arabic. Methodology.,pp.3614–3621.

Attia, M. and Tounsi, L. 2010. Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic In: Dublin: Technical report, The NCLT Seminar Series, DCU.

Attia, M., Tounsi, L., Pecina, P., van Genabith, J. and Toral, A. 2010. Automatic extraction of Arabic multiword expressions In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, pp. 19–27.

Attia, M.A. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar In: Advances in Natural Language Processing, Springer, Berlin, Heidelberg., pp. 87–98.

Attia, M.A. 2008. Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. PhD thesis, University of Manchester.

Atwell, E.S. 1988. Grammatical analysis of English by statistical pattern recognition In: Pattern Recognition. Springer, pp. 626–635.

Azmi, A.M. and Almajed, R.S. 2015. A survey of automatic Arabic diacritization techniques. Natural Language Engineering. 21(3),pp.477–495.

Badawi, E.S., Carter, M. and Gully, A. 2013. Modern written Arabic: A comprehensive grammar. Routledge.

Baldwin, T. 2005. Deep lexical acquisition of verb–particle constructions. Computer Speech & Language. 19(4),pp.398–414.

Baldwin, T. and Kim, S.N. 2010. Multiword expressions. Handbook of Natural Language Processing, second edition. Morgan and Claypool.

Baldwin, T. and Tanaka, T. 2004. Translation by machine of complex nominals: Getting it right In: Proceedings of the Workshop on Multiword Expressions: Integrating Processing. Association for Computational Linguistics, pp. 24–31.

Bannard, C. and Lieven, E. 2012. Formulaic language in L1 acquisition. Annual Review of Applied Linguistics. 32,pp.3–16.

Bar, K., Diab, M. and Hawwari, A. 2014. Arabic Multiword Expressions In: Language, Culture, Computation. Computational Linguistics and Linguistics. Springer, pp. 64–81.

Bartsch, S. 2004. Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Gunter Narr Verlag.

Basili, R., Pazienza, M.T. and Velardi., P. 1994. A ``not-so-shallow'' parser for collocational analysis. Proc. of the 15th COLING (COLING 1994).,pp.447–453.

Begoña Villada Moirón, M. 2004. Discarding Noise in an Automatically Acquired Lexicon of Support verb Constructions. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).,pp.1859–1862.

Bejcek, E., Hajic, J., Stranák, P., Uresová, Z., Bejček, E., Hajič, J., Straňák, P. and Urešová, Z. 2017. Extracting Verbal Multiword Data from Rich Treebank Annotation In: TLT., pp. 13–24.

Bejček, E., Stranak, P., Pecina, P., Bejcek, E., Stranák, P., Pecina, P., Bejček, E., Stranak, P. and Pecina, P. 2013. Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. Proceedings of the 9th Workshop on Multiword Expressions. (June),pp.106–115.

Benson, M., Benson, E. and Ilson, R. 1997. The BBI dictionary of English word combinations. John Benjamins Pub. Co.

Bergsma, S. and Wang, Q.I. 2007. Learning Noun Phrase Query Segmentation. In: EMNLP-CoNLL., pp. 819–826.

Berry-Rogghe, G. 1973. The computation of collocations and their relevance in lexical studies. The computer and literary studies.,pp.103–112.

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

Biber, D., Conrad, S., & Cortes, V. 2003. Lexical bundles in speech and writing: an initial taxonomy. In G. N. Leech, T. McEnery, A. Wilson & P. Rayson (Eds.), Corpus linguistics by the lune. New York, NY: Peter Lang.

Biber Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). Longman Grammar of Spoken and Written English. Harlow: Longman.

Biemann, C., Bontcheva, K., de Castilho, R.E., Gurevych, I. and Yimam, S.M. 2017. Collaborative Web-based Tools for Multi-layer Text Annotation In: Handbook of Linguistic Annotation. Springer, pp. 229–256.

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

Bollinger, L.C. 1976. Freedom of the press and public access: Toward a theory of partial regulation of the mass media. Michigan Law Review. 75(1),pp.1–42.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. and Shoul, M. 2010. Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts In: International Arab conference on information technology. Benghazi Libya, pp. 1–6.

Boulaknadel, S., Daille, B. and Aboutajdine, D. 2008. A Multi-Word Term Extraction Program for Arabic Language In: LREC., pp. 1485–1488.

Bounhas, I. and Slimani, Y. 2009. A hybrid approach for Arabic multi-word term extraction In: 2009 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, pp. 1–8.

Bourigault, D. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. Actes de COLING-92.,pp.977–981.

Bresnan, J., Kaplan, R.M., Peters, S. and Zaenen, A. 1982. Cross-serial dependencies in Dutch. Linguistic Inquiry. 13(4),pp.613–635.

Brezina, V., McEnery, T. and Wattam, S. 2015. Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics. 20(2),pp.139–173.

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G. and Shein, F. 2015. Building a Lexicon of Formulaic Language for Language Learners. Proc. of MWE.,pp.96–104.

Brooks, P.J. and Tomasello, M. 1999. How children constrain their argument structure constructions. Language.,pp.720–738.

Bruce, T. and Pedersen, R. 1996. What to Infer from a Description In: Dallas, TX.: Southern Methodist University.

Buckwalter, T. 2002. Buckwalter {Arabic} Morphological Analyzer Version 1.0.

Buckwalter, T. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02. ISBN 1-58563-324-0.

Burger, H. 2007. Phraseologie. Walter de Gruyter.

Burger, S., Constantini, E. and Pianesi, F., 2003. Communicative Strategies and Patterns of Multimodal Integration in a Speech-to-Speech Translation System. In Proceedings of MT Summit IX (pp. 32-39).

Burger, S. and Sloane, Z. 2004. The isl meeting corpus: Categorical features of communicative group interactions In: Proc. ICASSP-2004 Meeting Recognition Workshop.

Bussmann, H. 2006. Routledge dictionary of language and linguistics. Routledge.

Butt, M. 1999. A grammar writer's cookbook. CSLI Lecture Notes.

Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C. and Zampolli, A. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons In: LREC., pp. 1934–1940.

Cardey, S., Chan, R. and Greenfield, P. 2006. The development of a multilingual collocation dictionary In: Proceedings of the Workshop on Multilingual Language Resources and Interoperability. Association for Computational Linguistics, pp. 32–39.

Carpuat, M. and Diab, M. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. . (June),pp.242–245.

Chomsky, N. and Halle, M. 1965. Some controversial questions in phonological theory. Journal of linguistics. 1(2),pp.97–138.

Choueka, Y. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases In: RIAO 88:(Recherche d'Information Assistée par Ordinateur). Conference., pp. 609–623.

Church, K., Hanks, P., Hindle, D., Gale, W. (1991) "Using Statistics in Lexical Analysis," in Zernik (ed), Lexical Acquisition: Using On-line Resources to Build a Lexicon, Lawrence Erlbaum, pp. 115-164.

Church, K.W., Hanks, N.J.P. and Hanks, P. 1990. Word Association Norms, Mutual Information, and Lexicography. Computational linguistics. 16(1),pp.22–29.

Clark, E. V 2008. First Language Acquisition. Cambridge University Press.

Clear, J.H. 1993. The British national corpus In: The digital word. MIT Press, pp. 163–187.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20(1),pp.37–46.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. and Todirascu, A. 2017. Multiword expression processing: a survey. Computational Linguistics. (Just Accepted),pp.1–92.

Constant, M., Roux, J. Le and Sigogne, A. 2013. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. ACM Transactions on Speech and Language Processing. 10(3),pp.1–24.

Corpas Pastor, G. and Seghiri, M. 2010. Size matters: a quantative approach to corpus representativeness. Language, translation, reception. To honor Julio Cósar Santoyo. 1,pp.111–145.

Ruiz Costa-Jussà, M., Daudaravicius, V. and Banchs, R.E., 2010. Integration of statistical collocation segmentations in a phrase-based statistical machine translation system. In EAMT 2010: proceedings of the 14th annual conference of the European Association for Machine Translation.

Cowie, A.P. 1998. Phraseology: Theory, analysis, and applications. OUP Oxford.

Cowie, A.P. 2001. Speech formulae in English: problems of analysis and dictionary treatment. Groninger Arbeiten zur germanistischen Linguistik. (44),pp.1–12.

Cunningham, H. 2002. GATE, a general architecture for text engineering. Computers and the Humanities. 36(2),pp.223–254.

Czerepowicka, M. and Savary, A., 2015. SEJF-A Grammatical Lexicon of Polish Multiword Expressions. In Language and Technology Conference (pp. 59-73). Springer, Cham.

Daille, B. 1994. Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques. (Combined approach for terminology extraction : lexical statistics and linguistic filtering), PhD thesis, Paris.

Dale, R. 2010. Classical approaches to natural language processing. In Indurkhya, N. and Damerau, F.J. eds., 2010. Handbook of natural language processing (Vol. 2). CRC Press.

Daoud, D., Al-Kouz, A. and Daoud, M. 2016. Time-sensitive Arabic multiword expressions extraction from social networks. International Journal of Speech Technology. 19(2),pp.249–258.

Darwish, K. 2014. Arabic Information Retrieval. Foundations and Trends in Information Retrieval. 7(4),pp.239–342.

Darwish, K. and Mubarak, H. 2016. Farasa: A New Fast and Accurate Arabic Word Segmenter. In: The10th International Conference on Language Resources and Evaluation. Portorož, Slovenia: LREC.

Diab, M.T. and Krishna, M. 2009. Handling sparsity for verb noun MWE token classification In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. Association for Computational Linguistics, pp. 96–103.

Dias, G., Guilloré, S., Bassano, J.-C. and Pereira Lopes, J.G. 2000. Combining linguistics with statistics for multiword term extraction: A fruitful association? In: Content-Based Multimedia Information Access-Volume 2. HISDC, pp. 1473–1491.

Dickins, J., Hervey, S. and Higgins, I. 2016. Thinking Arabic translation: A course in translation method: Arabic to English. Taylor & Francis.

Dickins, J. and Watson, J.C.E. 1999. Standard Arabic Student's Book: An Advanced Course. Cambridge University Press.

Dipper, S., Götze, M. and Stede, M. 2004. Simple annotation tools for complex annotation tasks: an evaluation In: Proceedings of the LREC Workshop on XML-based richly annotated corpora., pp. 54–62.

Dorgeloh, H. and Wanner, A. 2009. Formulaic argumentation in scientific discourse. Formulaic language. 2,pp.523–544.

Dukes, K., Atwell, E. and Sharaf, A.-B.M. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank In: LREC. Citeseer, pp. 1822–1827.

Dukes, K. and Buckwalter, T. 2010. A dependency treebank of the Quran using traditional Arabic grammar In: Informatics and Systems (INFOS), 2010 The 7th International Conference on. IEEE, pp. 1–7.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics. 19(1),pp.61–74.

Durrant, P.L. 2008. High frequency collocations and second language learning. PhD thesis, University of Nottingham.

EbdAlrzAq, E. 2007. AlmtlAzmAt AllfZyp fy Allgp wAlqwAmys AlErbyp. 'Multiword expressions in Arabic dictionaries' Alatrash Publisher, Tunis.

El-haj, M. and Rayson, P. 2016. OSMAN – A Novel Arabic Readability Metric. In: The10th International Conference on Language Resources and Evaluation. Portorož, Slovenia: LREC.

El-Haj, M., Rayson, P., Piao, S. and Wattam, S. 2017. Creating and Validating Multilingual Semantic Representations for Six Languages : Expert versus Non-Expert Crowds In: Association for Computational Linguistics, pp. 61–71.

Elewa, A.-H. 2004. Collocation and synonymy in classical Arabic: A corpus-based study. PhD thesis, University of Manchester.

Ellis, N.C. 1996. Working memory in the acquisition of vocabulary and syntax: Putting language in good order. The Quarterly Journal of Experimental Psychology: Section A. 49(1),pp.234–250.

Ellis, N.C., Simpson-vlach, R. and Maynard, C. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. Tesol Quarterly. 42(3),pp.375–396.

Erman, B. and Warren, B., 2000. The idiom principle and the open choice principle. Text-Interdisciplinary Journal for the Study of Discourse, 20(1), pp.29-62.

Evert, S. 2008. Corpora and collocations. Corpus Linguistics. An International Handbook. 2, pp.1212–1248.

Evert, S. 2004. The Statistics of Word Co-occurrences Word Pairs and Collocations. PhD thesis, Institut fur maschinelle Sprachverarbeitung Universitat Stuttgart.

Evert, S., Heid, U. and Spranger, K. 2004. Identifying Morphosyntactic Preferences in Collocations. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon (Portugal), pp. 907–910.

Evert, S. and Kermes, H. 2003. Experiments on Candidate Data for Collocation Extraction. Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03).,pp.83–86.

Evert, S. and Kermes, H. 2002. The influence of linguistic pre-processing on candidate data In: Proceedings of the Workshop on Computational Approaches to Collocations. Citeseer.

Evert, S. and Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01.,pp.188–195.

Evert, S. and Krenn, B. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language. 19(4),pp.450–466.

Farahmand, M., Smith, A. and Nivre, J. 2015. A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds. Proceedings of NAACL-HLT 2015.,pp.29–33.

Farghaly, A. 1987. Three level morphology for Arabic In: Proceedings of the Arabic Morphology Workshop (AMW'87). Linguistic summer institute, Stanford CA

Farghaly, A. and Shaalan, K. 2009. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP). 8(4)

Fehri, A.F. 2012. Key features and parameters in Arabic grammar. John Benjamins Publishing.

Fellbaum, C., Geyken, A., Herold, A., Koerner, F. and Neumann, G. 2006. Corpus-based studies of German idioms and light verbs. International Journal of Lexicography. 19(4), pp.349–360.

Fillmore, C.J. 1979. On fluency In: Individual differences in language ability and language behavior. Elsevier, pp. 85–101.

Fillmore, C.J., Kay, P. and O'connor, M.C. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. Language.,pp.501–538.

Firth, J. 1951. Papers in Linguistic [s J]. Oxford University Press.

Firth, J.R. 1961. Papers in Linguistics 1934-1951: Repr. Oxford University Press.

Firth, R. 1957. 2. A Note on Descent Groups in Polynesia. Man. 57,pp.4–8.

Fort, K. 2016. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects. John Wiley & Sons.

Francopoulo, G. 2013. LMF lexical markup framework. Hoboken, NJ; London: ISTE Ltd.

Frantzi, K.T. and Ananiadou, S. 1996. Extracting nested collocations. Proceedings of the 16th conference on Computational linguistics. 1,pp.41–46.

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P. and Piao, S. 2001. The METER corpus: a corpus for analysing journalistic text reuse In: Proceedings of the corpus linguistics 2001 conference., pp. 214–223.

Gardner, D. and Davies, M. 2014. A new academic vocabulary list. Applied Linguistics. 35(3),pp.305–327.

Garrao, M., Quental, V., Caminada, N. and Bick, E. 2008. The Identification and Description of Frozen Prepositional Phrases through a Corpus-Oriented Study In: Computational Processing of the Portuguese Language. Springer, pp. 220–223.

Garside, R. 1987. The CLAWS word-tagging system.

Garside, R. and Rayson, P. 1997. Higher-level annotation tools. Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London.,p.179–193.

Garside, R., Sampson, G. and Leech, G. 1988. The computational analysis of English: A corpus-based approach. Longman.

Giacomini, L. 2017. Designing a Learner's Dictionary with Phraseological Disambiguators. . 10596,pp.290–305.

Goldman, J.-P., Nerima, L. and Wehrli, E. 2001. Collocation extraction using a syntactic parser In: Proceedings of the ACL Workshop on Collocations., pp. 61–66.

Gralinski, F., Savary, A., Czerepowicka, M. and Makowiecki, F. 2010. Computational Lexicography of Multi-Word Units: How Efficient Can It Be? In: Workshop Multiword Expressions: from Theory to Applications., p. .

Granger, S. and Meunier, F. 2008. Phraseology: an interdisciplinary perspective. Philadelphia: John Benjamins Pub. Company.

Granger, S. and Paquot, M. 2008. Disentangling the phraseological web. Phraseology. An interdisciplinary perspective. (Gross),pp.27–50.

Granger, S. and Paquot, M. 2012. Electronic lexicography. Oxford University Press.

Grégoire, N.H.W. 2009. Untangling Multiword Expressions, A study on the representation and variation of Dutch multiword expressions. LOT.

Gries, S.T. 2008. Phraseology and linguistic theory: A brief survey. Phraseology: An interdisciplinary perspective.,pp.3–25.

Guenthner, F. and Blanco, X. 2004. Multi-lexemic expressions: an overview.

Gurevych, I., Eckle-Kohler, J. and Matuschek, M. 2016. Linked Lexical Knowledge Bases: Foundations and Applications.

Gurrutxaga, A. and Alegria, I. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. (June),pp.2–7.

Habash, N., Faraj, R. and Roth, R. 2009. Syntactic annotation in the Columbia Arabic treebank In: Proceedings of MEDAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt., pp. 125–132.

Habash, N. and Rambow, O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05. (June),pp.573–580.

Habash, N.Y. 2010. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies. 3(1),pp.1–187.

Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., Beška, E., Hajič, J., Smrz, O., Zemánek, P., Šnaidauf, J. and Beška, E. 2004. Prague Arabic dependency treebank: Development in data and tools In: Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools., pp. 110–117.

Hanks, P. 2013. Lexical Analysis. The MIT Press.

Hawwari, A., Attia, M. and Diab, M. 2014. A Framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic. ANLP 2014.,p.48.

Hawwari, A., Bar, K. and Diab, M. 2012. Building an Arabic multiword expressions repository In: Proceedings of the ACL 2012 joint workshop on statistical parsing and semantic processing of morphologically rich languages, Jeju. Association for Computational Linguistics. Citeseer, pp. 24–29.

Hearst, M. a. and Hearst, M. a. 1998. Automated discovery of wordnet relations. WordNet: an electronic lexical database.,pp.131–153.

Hearst, M.A. 1992. Automatic acquisition of hyponyms from large text corpora In: Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, pp. 539–545.

Heid, U. 1998. A linguistic bootstrapping approach to the extraction of term candidates from German text. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. 5(2),pp.161–181.

Ho-Dac, L.-M. 2009. Introduction to Linguistic Annotation and Text Analytics.

Holes, C. 2004. Modern Arabic: Structures, functions, and varieties. Georgetown University Press.

Huang, R.E.N., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A. and Prévot, L. (Ed. . 2010. Ontology and the Lexicon: a natural language processing perspective (Studies in Natural Language Processing).

Hunston, S. and Francis, G. 2000. Pattern grammar: A corpus-driven approach to the lexical grammar of English. Computational Linguistics. 27(2),pp.318–320.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. English for specific purposes. 27(1),pp.4–21.

Ide, N. and Pustejovsky, J. 2017. Handbook of Linguistic Annotation. Springer.

Induction, V., Patterns, C., Relations, S. and Seretan, V. 2005. Relations. ,pp.1698–1699.

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

Jurafsky, D. and Martin, J.H. 2007. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Computational Linguistics. 26(4),pp.638–641.

Justeson, J.S. and Katz, S.M. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering. 1(1),pp.9–27.

Kato, A., Shindo, H. and Matsumoto, Y. 2013. Construction of an English Dependency Corpus incorporating Compound Function Words. . (Section 4),pp.1667–1671.

Khemakhem, A., Gargouri, B., Haddar, K. and Ben Hamadou, A. 2013. LMF for Arabic. LMF Lexical Markup Framework.,pp.83–98.

Kilgarriff, A. 2005. Language is never, ever, ever, random. Corpus linguistics and linguistic theory. 1(2),pp.263–276.

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

Kilgarriff, A. and Tugwell, D. 2001. Word sketch-extraction and display of significant collocations for lexicography.pdf.

Kim, S., Yoon, J. and Song, M. 2001. Automatic extraction of collocations from Korean text. Computers and the Humanities. 35(3),pp.273–297.

Korkontzelos, I. 2010. Unsupervised Learning of Multiword Expressions. . (September).

Korkontzelos, I. and Manandhar, S. 2010. Can recognising multiword expressions improve shallow parsing? In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 636–644.

Kremmel, B., Brunfaut, T. and Alderson, J.C. 2015. Exploring the Role of Phraseological Knowledge in Foreign Language Reading. Applied Linguistics. (January),p.amv070.

Krenn, B. 2000a. CDB - A Database of Lexical Collocations. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00).

Krenn, B. 2008. Description of evaluation resource–German PP-verb data In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Citeseer, pp. 7–10.

Krenn, B. 2000b. The usual suspects: Data-oriented models for identification and representation of lexical collocations.

Krenn, B. and Evert, S. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. Proceedings of the ACL Workshop on Collocations.,pp.39–46.

Krenn, B., Evert, S. and Zinsmeister, H. 2004. Determining intercoder agreement for a collocation identification task. Proceedings of KONVENS.,pp.89–96.

Kuiper, K., McCann, H., Quinn, H., Aitchison, T. and van der Veer, K. 2003. SAID: A syntactically annotated idiom dataset. Linguistic Data Consortium, LDC2003T10, Pennsylvania.

Lambert, P. and Banchs, R. 2005. Data inferred multi-word expressions for statistical machine translation. Proceedings of Machine Translation Summit X.,pp.396–403.

Langacker, R.W. 1991. Cognitive grammar. Linguistic theory and grammatical description.,pp.275–306.

Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J., 2014. Automated grammatical error detection for language learners. Synthesis lectures on human language technologies, 7(1), pp.1-170.

Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. 2010. Automated grammatical error detection for language learners. Synthesis lectures on human language technologies. 3(1),pp.1–134.

Leech, G. 1993. 100 million words of English. English Today. 9(01),p.9.

Leech, G., Garside, R. and Atwell, E.S. 1983. The automatic grammatical tagging of the LOB corpus. ICAME Journal: International Computer Archive of Modern and Medieval English Journal. 7,pp.13–33.

Leech, G.N. 1997. Introducing corpus annotation. Corpus Annotation: Linguistic Information from Computer Text Corpora.,pp.1–18.

Leech, G.N., Rayson, P. and Wilson, A. 2001. Word frequencies in written and spoken English : based on the British National Corpus. Harlow: Longman.

Lelubre, X. 2001. A Scientific Arabic Terms Data Base: Linguistic Approach for a Representation of Lexical and Terminological Features In: ACL 39th Annual Meeting. Citeseer, pp. 66–72.

Leonardo Zilio Luiz Henrique Longhi Rossi, Rafael Martins Feitosa, L.S., Zilio, L., Svoboda, L., Rossi, L.H.L. and Feitosa, R.M. 2011. Automatic extraction and evaluation of MWE In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. IEEE, pp. 214–218.

Lewis, M. and Conzett, J. 2000. Teaching collocation: Further developments in the lexical approach. Language Teaching Publications Hove.

Lewis, M. and Gough, C. 1997. Implementing the lexical approach: Putting theory into practice. Language Teaching Publications Hove.

Li, W., Zhang, X., Niu, C., Jiang, Y. and Srihari, R. 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.

Lin, D. 1999. Automatic identification of non-compositional phrases. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -.,pp.317–324.

Lin, D. 1998. Extracting collocations from text corpora In: First workshop on computational terminology. Citeseer, pp. 57–63.

Löfberg, L., Piao, S., Rayson, P., Juntunen, J.-P., Nykänen, A. and Varantola, K. 2005. A semantic tagger for the {Finnish} language. Proceedings of the Corpus Linguistics 2005 conference.

Losnegaard, G.S., Sangati, F., Escartín, C.P., Savary, A., Bargmann, S. and Monti, J. 2016. PARSEME Survey on MWE Resources In: The tenth International Conference on Language Resources and Evaluation (LREC 2016).

Lu, X. 2014. Computational methods for corpus annotation and analysis. New York; London: Springer.

Luong, M.-T., Pham, H. and Manning, C.D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In: the proceedings of NAACL pp.11-19.

Maamouri, M. and Bies, A. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages. Association for Computational Linguistics, pp. 2–9.

Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus In: NEMLAR conference on Arabic language resources and tools. Cairo, pp. 466–467.

Maamouri, M., Bies, A., Krouna, S., Gaddeche, F. and Bouziri, B. 2009. Penn Arabic treebank guidelines. Linguistic Data Consortium.

Manning, Christopher; Schütze, H. 1999. Collocations. Foundations of Statistical Natural Language Processing.,pp.151–189.

Manning, C.D. and Schütze, H. 1999. Foundations of statistical natural language processing. MIT Press.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D. 2014. The stanford corenlp natural language processing toolkit In: ACL (System Demonstrations)., pp. 55–60.

Marcus, M.P., Marcinkiewicz, M.A., Santorini, B. and Marcinkiewicz, M.A. 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics. 19(2),pp.313–330.

Martinez, R. 2011. The development of a corpus-informed list of formulaic sequences for language pedagogy. PhD thesis, University of Nottingham.

Marton, Y., Habash, N. and Rambow, O. 2013. Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. Computational Linguistics. 39(1),pp.161–194.

Maynard, D., Cunningham, H. and Bontcheva, K. 2004. Automatic Language-Independent Induction of Gazetteer Lists. In Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp.713-716

McCarthy, D., Keller, B. and Carroll, J. 2003. Detecting a continuuum of compositionality in phrasal verbs. In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. 213,pp.73–80.

McCarthy, J.J. 1981. A prosodic theory of nonconcatenative morphology. Linguistic inquiry. 12(3),pp.373–418.

McCarthy, J.J. and Prince, A. 1994. Prosodic Morphology I: Constraint Interaction and Satisfaction. Yearbook of morphology 1993. (January),pp.79–153.

McEnery, T. and Gabrielatos, C. 2008. English Corpus Linguistics. The Handbook of English Linguistics.,pp.33–71.

McEnery, T. and Hardie, A. 2011. Corpus linguistics: Method, theory and practice. Cambridge University Press.

McEnery, T., Langé, J.-M., Oakes, M. and Véronis, J. 1997. The exploitation of multilingual annotated corpora for term extraction. Corpus annotation---linguistic information from computer text corpora.,pp.220–230.

Mcinnes, B.T. 2004. Extending the Log Likelihood Measure to Improve Collocation Identification. PhD thesis, University of Minnesota.

Meghawry, S., Elkorany, A., Salah, A. and Elghazaly, T. 2015. Semantic Extraction of Arabic Multiword Expressions. Computer Science & Information Technology. 5(2),pp.21–31.

Mel'cuk, I. 2003. Collocations: définition, rôle et utilité. Travaux et recherches en linguistique appliquée. Série E, Lexicologie et lexicographie. (1),pp.23–31.

Mel'cuk, I. 1995a. Phrasemes in language and phraseology in linguistics. Idioms: Structural and psychological perspectives.,pp.167–232.

Mel'čuk, I., 1998. Collocations and lexical functions. Phraseology. Theory, analysis, and applications, pp.23-53.

Mel'Čuk, I. 2012. Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology. 3(1),pp.31–56.

Mel'čuk, I. a 1998. Collocations and lexical functions. Phraseology. Theory, analysis, and applications. (1),pp.23–53.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. 2013. Distributed representations of words and phrases and their compositionality In: Advances in neural information processing systems., pp. 3111–3119.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. 1990. Introduction to wordnet: An on-line lexical database. International journal of lexicography. 3(4),pp.235–244.

Moirón, M.B.V. 2005. Data-driven identification of fixed expressions and their modifiability. PhD thesis, University of Groningen

Monti, J. 2015. Multi-word unit processing in machine translation. Developing and using language resources for multi-word unit processing in machine translation. PhD thesis, University of Salerno.

Moon, R. 1998. Fixed expressions and idioms in English: A corpus-based approach. Oxford University Press.

Najar, D., Mesfar, S. and Ghezela, H. Ben 2015. A large terminological dictionary of Arabic compound words In: International NooJ Conference. Springer, pp. 16–28.

Najar, D., Mesfar, S. and Ghezela, H. Ben 2016. A large terminological dictionary of Arabic compound words In: Communications in Computer and Information Science. Springer, pp. 16–28.

Nakagawa, H. and Mori, T. 2003. Automatic term recognition based on statistics of compound nouns and their components. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. 9(2),pp.201–219.

Nattinger, J.R. and DeCarrico, J.S. 1992. Lexical phrases and language teaching. Oxford: Oxford University Press.

Nesselhauf, N. 2005. Collocations in a learner corpus. John Benjamins Publishing.

Castagnoli, S., 2014. Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology. In Proceedings of the 10th Workshop on Multiword Expressions (MWE) (pp. 57-61).

Obeid, O., Bouamor, H., Zaghouani, W., Ghoneim, M., Hawwari, A., Alqahtani, S., Diab, M. and Oflazer, K. 2016. Mandiac: A web-based annotation system for manual arabic diacritization In: The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media.

Odijk, J. 2013a. DUELME: Dutch electronic lexicon of multiword expressions. LMF Lexical Markup Framework.,pp.133–144.

Odijk, J. 2013b. Identification and lexical representation of multiword expressions In: Essential Speech and Language Technology for Dutch. Springer, pp. 201–217.

Odijk, J. 2003. Towards a standard for multi-word expressions. ISLE Project Report, February.

Ohlrogge, A. 2009. Formulaic expressions in intermediate EFL writing assessment. Formulaic language. 2,pp.387–404.

Pal, S., Naskar, S. and Bandyopadhyay, S., 2013. A hybrid word alignment model for phrase-based statistical machine translation. In Proceedings of the Second Workshop on Hybrid Approaches to Translation (pp. 94-101).

Pal, S., Naskar, S.K. and Bandyopadhyay, S. 2013b. MWE alignment in phrase based statistical machine translation. The XIV Machine Translation Summit.,pp.61–68.

Pal, S., Naskar, S.K., Pecina, P., Bandyopadhyay, S. and Way, A. 2010. Handling named entities and compound verbs in phrase-based statistical machine translation In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications., pp. 46–54.

Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic In: LREC., pp. 1094–1101.

Pawley, A., &F. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. J. Richards&R. Schmidt (eds.). Language and Communication, pp.191-226.

Pearce, D. 2002. A Comparative Evaluation of Collocation Extraction Techniques. Proceedings of the 3rd Edition of the Language, Resources and Evaluation Conference (LREC 2002).,pp.1530–1536.

Pecina, P. 2008a. A machine learning approach to multiword expression extraction In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Citeseer, pp. 54–61.

Pecina, P., 2005, June. An extensive empirical study of collocation extraction methods. In Proceedings of the ACL Student Research Workshop (pp. 13-18). Association for Computational Linguistics.

Pecina, P. 2009. Lexical Association Measures. PhD thesis, Charles University.

Pecina, P. 2010. Lexical association measures and collocation extraction. Language resources and evaluation. 44(1–2),pp.137–158.

Pecina, P. 2008b. Reference data for Czech collocation extraction In: Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)., pp. 11–14.

Pedersen, T., Banerjee, S., McInnes, B.T., Kohli, S., Joshi, M. and Liu, Y. 2011. The Ngram statistics package (text:: nsp): A flexible tool for identifying ngrams, collocations, and word associations In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. Association for Computational Linguistics, pp. 131–133.

Piao Bianchi, F., Dayrell, C., D'Egidio, A. and Rayson, P., S. and Piao Bianchi Dayrell, C., D'Egidio, A. and Rayson, P., S, F. 2015. Development of the multilingual semantic annotation system. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies.,pp.1268–1274.

Piao, S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, A.M. and Wilson, A. 2005. A large semantic lexicon for corpus annotation. Corpus Linguistics Conference 2005.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., Jiménez, R., Knight, D., Křen, M., Löfberg, L., Nawab, M.A., Shafi, J., Teh, P.L. and Mudraya, O. 2016. Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages In: Tenth International Conference on Language Resources and Evaluation. LREC, pp. 2614–2619.

Piao, S., Rayson, P., Watkins, G., Knight, D. and Donnelly, K., 2017, July. Towards a Welsh semantic tagger: creating lexicons for a resource poor language. In CL2017 conference, University of Birmingham, Birmingham (pp. 24-28).

Piao, S., Rayson, P., Mudraya, O., Wilson, A. and Garside, R. 2006. Measuring MWE Compositionality Using Semantic Annotation. Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. (July),pp.2–11.

Piao, S.S., Rayson, P., Archer, D. and McEnery, T. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. Computer Speech & Language. 19(4),pp.378–397.

Piao, S.S.L., Rayson, P., Archer, D., Wilson, A. and McEnery, T. 2003. Extracting multiword expressions with a semantic tagger In: Proceedings of the ACL 2003 workshop on Multiword expressions analysis, acquisition and treatment, Morristown, NJ, USA: Association for Computational Linguistics, pp. 49–56.

Ramisch, C. 2012. A generic and open framework for multiword expressions treatment: from acquisition to applications. Proceedings of ACL 2012 Student Research Workshop, pp.61–66.

Ramisch, C. 2015a. Multiword Expressions Acquisition: A Generic and Open Framework. Springer. London.

Ramisch, C. 2015b. State of the Art in MWE Processing In: Multiword Expressions Acquisition. Springer, pp. 53–102.

Ramisch, C., De Araujo, V. and Villavicencio, A. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions In: Proceedings of ACL 2012 Student Research Workshop. Association for Computational Linguistics, pp. 1–6.

Rayson, P. 2008. From key words to key semantic domains. International Journal of Corpus Linguistics. 13(4),pp.519–549.

Rayson, P. 2002. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis, University of Lancaster.

Rayson, P., Archer, D., Piao, S. and McEnery, T. 2004. The UCREL semantic analysis system. Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop.,pp.7–12.

Rayson, P., Piao, S., Sharoff, S., Evert, S. and Moirón, B.V. 2010. Multiword expressions: hard going or plain sailing? Language Resources and Evaluation. 44(1–2),pp.1–5.

Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y. 2009. Improving statistical machine translation using domain bilingual multiword expressions In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Association for Computational Linguistics, pp. 47–54.

Riedl, M. and Biemann, C. 2015. A Single Word is not Enough: Ranking Multiword Expressions Using Distributional Semantics In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 2430–2440.

Riedl, M. and Biemann, C., 2016. Impact of MWE resources on multiword recognition. In Proceedings of the 12th Workshop on Multiword Expressions (pp. 107-111).

Rikters, M. and Bojar, O. 2017. Paying Attention to Multi-Word Expressions in Neural Machine Translation. arXiv preprint arXiv:1710.06313.

Romary, L., Ide, N. and Kilgarriff, A. 2000. A formal model of dictionary structure and content In: Proceedings of Euralex 2000. Stuttgart, pp. 113–126.

Rosenhouse, Y. and Versteegh, K. 2006. Encyclopaedia of Arabic language and linguistics. Volume I, Brill Leiden, Boston.

Roth, R., Rambow, O., Habash, N., Diab, M. and Rudin, C. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pp. 117–120.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R. and Scheffczyk, J., 2016. FrameNet II: Extended theory and practice. Institut für Deutsche Sprache, Bibliothek.

Ryding, K.C. 2005. A reference grammar of modern standard Arabic. Cambridge university press.

Saad, M.K. and Ashour, W. 2010. Arabic morphological tools for text mining. Corpora. 18,p.19.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP In: Computational Linguistics and Intelligent Text Processing. Springer, pp. 1–15.

Saif, A. and Aziz, M. 2011. An automatic collocation extraction from Arabic corpus. Journal of Computer Science. 7(1),pp.6–11.

Saif, A.M., Aziz, M.J.A. and Saif 2011. An automatic collocation extraction from Arabic corpus. Journal of Computer Science. 7(1),pp.6–11.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B. and Losnegaard, G.S. 2015. PARSEME–PARSing and Multiword Expressions within a European multilingual network In: 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015).

Sawalha, M.S.S. 2011. Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora. PhD thesis, University of Leeds.

Schmitt, N. 2004. Formulaic sequences: Acquisition, processing, and use. John Benjamins Publishing.

Schmitt, N. 2010. Researching vocabulary: a vocabulary research manual. Basingstoke: Palgrave Macmillan.

Schmitt, N. and Martinez, R. 2012. A Phrasal Expressions List. Applied Linguistics. 33(3),p.299.

Schneider, N. 2014. Lexical Semantic Analysis in Natural Language Text. PhD thesis, Carnegie Mellon University.

Schneider, N., Mohit, B., Dyer, C., Oflazer, K. and Smith, N.A. 2013. Supersense Tagging for Arabic: the MT-in-the-Middle Attack In: HLT-NAACL. Citeseer, pp. 661–667.

Seretan, V. 2011. Syntax-based collocation extraction. Springer, London.

Seretan, V., Nerima, L. and Wehrli, E. 2004. Multi-word collocation extraction by syntactic composition of collocation bigrams In: Recent Advances in Natural Language Processing III., pp. 91–100.

Shahrour, A., Khalifa, S. and Habash, N. 2015. Improving Arabic diacritization through syntactic analysis In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing., pp. 1309–1315.

Shin, D. and Nation, P. 2008. Beyond single words: The most frequent collocations in spoken English. ELT Journal. 62(4),pp.339–348.

Shudo, K., Kurahone, A. and Tanabe, T. 2011. A comprehensive dictionary of multiword expressions In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 161–170.

Sidtis, D.V.L. 2011. Formulaic Expressions in Mind and Brain In: The Handbook of Psycholinguistic and Cognitive Processes. Routledge.

Silva, J. and Lopes, G., 2010, August. Towards automatic building of document keywords. In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1149-1157.

Da Silva, J.F., Dias, G., Guilloré, S. and Lopes, J.G.P. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units In: Portuguese Conference on Artificial Intelligence. Springer, pp. 113–132.

Sinclair, J. 1991. Corpus, concordance, collocation. Oxford University Press.

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.

Sinclair, J. and Renouf, A. 1988. A lexical syllabus for language learning. Vocabulary and Language Teaching.,pp.140–158.

Siyanova-Chanturia, A., Conklin, K. and 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),pp.251–272.

Smadja, F. 1993. Retrieving collocations from text: Xtract. Computational linguistics. 19(1),pp.143–177.

Smadja, F., McKeown, K.R. and Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: A statistical approach. Computational linguistics. 22(1),pp.1–38.

Smrž, O. 2007. ElixirFM In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages Common Issues and Resources, Semitic '07. Morristown, NJ, USA: Association for Computational Linguistics, p. 1-8.

Smrz, O. and Bielický, V. 2010. ElixirFM. Functional Arabic Morphology. PhD thesis, Charles University.

Stevens, M.E. and Giuliano, V.E. 1965. Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, Washington, 1964. US Government Printing Office.

Tan, L. and Pal, S. 2014. Manawi: Using Multi-Word Expressions and Named Entities to Improve Machine Translation. Proceedings of the Ninth Workshop on Statistical Machine Translation. (2010),pp.201–206.

Tanabe, T., Takahashi, M. and Shudo, K. 2014. A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. Computer Speech and Language. 28(6),pp.1317–1339.

Taslimipoor, S., Fazly, A. and Hamzeh, A. 2012. A class-based acceptability measure for persian light verb constructions. AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing.,pp.250–255.

Thanopoulos, A., Fakotakis, N. and Kokkinakis, G. 2002. Comparative Evaluation of Collocation Extraction Metrics. In Proceedings of the 3rd Language Resources Evaluation Conference.,pp.620–625.

Todiraşcu, A., Tufiş, D., Heid, U., Gledhill, C., Ştefanescu, D., Weller, M. and Rousselot, F. 2008. A hybrid approach to extracting and classifying verb+ noun constructions. Proceedings of the 6th International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Tschichold, C. Multi-word units in natural language processing. PhD thesis, Basel University.

Hatier, S., Augustyn, M., Tran, T.T.H., Yan, R., Tutin, A. and Jacques, M.P., 2016. French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In Proceedings of Euralex (pp. 355-366).

Viera, A.J. and Garrett, J.M. 2005. Understanding interobserver agreement: The kappa statistic. Family Medicine. 37(5),pp.360–363.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M. and Ramisch, C. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering In: EMNLP-CoNLL. Citeseer, pp. 1034–1043.

Vincze, V., Nagy, T.I. and Berend, G., 2011a, June. Detecting noun compounds and light verb constructions: a contrastive study. In Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World Association for Computational Linguistics, pp. 116-121.

Vincze, V., Nagy, I. and Berend, G. 2011b. Multiword expressions and named entities in the Wiki50 corpus In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011., pp. 289–295.

Vintar, Š., Vintar, Š., Fišer, D. and Fišer, D. 2008. Harvesting Multi-Word Expressions from Parallel Corpora. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). (Fišer),pp.1091–1096.

Wehrli, E., Seretan, V. and Nerima, L. 2010. Sentence analysis and collocation identification In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications., pp. 28–36.

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

Van der Wouden, T. 2001. Collocational behaviour in non-content words In: ACL/EACL Workshop on Collocations, Toulouse, France.pp.16-23.

Wray, A. 2013. Formulaic language. Language Teaching. 46(03),pp.316–334.

Wray, A. 2002a. Formulaic language and the lexicon. Cambridge University Press Cambridge.

Wray, A. 2002b. Formulaic Language in Computer-supported Communication: Theory Meets Reality. Language Awareness. 11(2),pp.114–131.

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,pp.231–254.

Wray, A. and Perkins, M.R. 2000. The Functions of Formulaic Language: An Integrated Model. Language & Communication. 20(1),pp.1–28.

Wray, A. 2009. Identifying formulaic language: Persistent challenges and new opportunities. *Formulaic language*. **1**,pp.27–51.

Wu, H. and Zhou, M. 2003. Synonymous collocation extraction using translation information. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03. 1(49),pp.120–127.

Wulff, S., Swales, J.M. and Keller, K. 2009. "We have about seven minutes for questions": The discussion sessions from a specialized conference. English for specific purposes. 28(2),pp.79–92.

Yazdani, M., Farahmand, M. and Henderson, J. 2015. Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. (September),pp.1733–1742.

Zaninello, A. and Nissim, M. 2010. Creation of Lexical Resources for a Characterisation of Multiword Expressions in Italian. Proc. of the Seventh LREC (LREC 2010)., pp.654–661.

Zarrieß, S. and Kuhn, J. 2009. Exploiting translational correspondences for pattern-independent MWE identification In: Proceedings of the Workshop on Multiword Expressions Identification, Interpretation, Disambiguation and Applications - MWE '09. Morristown, NJ, USA: Association for Computational Linguistics, pp. 23–30.

APPENDIX A. THE GERMAN STANDARD DIN 31636 FOR RENDERING ROMANIZED ARABIC

Nu	Arabic letters	Nu	Arabic letters
1	أ ’	18	ع ‘
2	ب b	19	غ ġ
3	ت t	20	ف f
4	ث ṯ	21	ق q
5	ج ǧ	22	ك k
6	ح ḥ	23	ل l
7	خ ḫ	24	م m
8	د d	25	ن n
9	ذ d	26	هـ h
10	ر r	27	و w
11	ز z	28	ي y
12	س s	29	◌َ (short vowel) a
13	ش š	30	◌ُ (short vowel) u
14	ص ṣ	31	◌ِ (short vowel) i
15	ض ḍ	32	ا (long vowel) ā
16	ط ṭ	33	و (long vowel) ū
17	ظ ẓ		

APPENDIX B. LIST OF MWE TERMS AND DEFINITIONS

References	Definitions
Firth (1957, 181)	“Collocations of a given word are statements of the habitual and customary places of that word”
Firth (1968, 182)	“Collocations are actual words in habitual company”
Weinreich (1967,42)	"[...] any expression in which at least one constituent is polysemous, and in which a selection of a subsense is determined by the verbal context, [is called] a *phraseological unit*. A phraseological unit that involves at least two polysemous constituents, and in which there is a reciprocal contextual selection of subsenses, will be called an *idiom*."
Fraser (1970,22)	"I shall regard an idiom as a constituent or series of constituents for which the semantic interpretation is not a compositional function of the formatives of which it is composed."
Cowie (1978, 132)	“the co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern”
Hausmann (1985)	“typical, specific and characteristic combination of two words”
Cruse (1986, 40)	“The term collocation will be used to refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent”
Kjellmer (1987, 133)	“a sequence of words that occurs more than once in identical form (. . .) and which is grammatically well structured”
Choueka (1988)	“a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components”
Benson (1990:131)	“A collocation is an arbitrary and recurrent word combination”
Sinclair (1991, 170)	“Collocation is the co-occurrence of two or more words within a short space of each other in a text”
Fontenelle (1992, 222)	“The term collocation refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure”
Smadja (1993, 143)	“recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages”
Schenk (1994,2)	"Idioms are expressions for which a literal interpretation does not yield the correct meaning of the idiomatic expression."

Mel'ˇcuk (1995,167)	"An idiom is a multi-lexemic expression E whose meaning cannot be deduced by the general rules of the language in question from the meaning of the constituent lexemes of E, their semantically loaded morphological characteristics (if any) and their syntactic configuration."
Vander Wouden (1997, 5)	"Collocation: idiosyncratic restriction on the combinability of lexical items"
O'Grady (1998,279)	"I assume that idioms have a meaning that is not a simple function of the literal (i.e., non-figurative) meaning of their parts and that they manifest a high degree of conventionality in the choice of component lexical items."
Manning and Schütze (1999, 151)	"A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things"
Riehemann (2001,2)	"I use the term 'idiom' to refer to an expression made up out of two or more words, at least one of which does not have any of the meanings it can have outside of the expression. As will become clear from the discussion below, this is not intended as an exact definition."
McKeown and Radev (2000, 507)	"Collocations (. . .) cover word pairs and phrases that are commonly used in language, but for which no general syntactic and semantic rules apply".
Polguère (2000, 518)	"The notion of collocation refers to semi-idiomatic expressions L1+L2 such that one of the components, the collocate, is chosen to express a given meaning, in a specific syntactic role, contingent upon the choice of the other component, called the base of the collocation"
Lea and Runcie (2002, vii)	"Collocation is the way words combine in a language to produce natural-sounding speech and writing"
Sag et al. (2002, 7)	"Institutionalized phrases are semantically and syntactically compositional, but statistically idiosyncratic. (. . .) We reserve the term *collocation* to refer to any statistically significant co-occurrence, including all forms of MWE (. . .) and compositional phrases".
Evert (2004b, 17)	"A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon".
Bartsch (2004, 76)	"lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other"
Krenn (2008, 7)	"Collocations in our terms are lexically motivated word combinations that constitute"

APPENDIX C. COMPLETE NOTATION OF STANFORD ARABIC PARSER

The tagset of SAP:

Tag code	Explanation
(DT)?NN.*	noun
VB.*	verb
(DT)?JJ.*	adjective
W?RB	adverb
CC	conjunction
IN	prepostion
PRP.?	pronoun
CD	cardinal number
ADJ	adj
CC	Coordinating conjunction
CD	Cardinal number
DT	determiner
DTJJ	adjective with the determiner “Al” (ال)
DTJJR	adjective, comparative with the determiner “Al” (ال)
DTNN	noun, singular or mass with the determiner “Al” (ال)
DTNNP	Proper noun, singular with the determiner “Al” (ال)
DTNNPS	Proper noun, plural with the determiner “Al” (ال)
DTNNS	noun, plural with the determiner “Al” (ال)
IN	Preposition or subordinating conjunction
JJ	adjective
JJR	Adjective, comparative
NN	noun, singular or mass

NNP	Proper noun, singular
NNPS	Proper noun, plural
NNS	noun, plural
NOUN	noun
PRP	Personal pronoun
PRP$	Possessive pronoun
PUNC	punctuation
RB	adverb
RP	particle
UH	interjection
VB	verb, base form
VBD	Verb, past tense
VBG	verb, gerund or present participle
VBN	verb, past participle
VBP	Verb, non-3rd person singular present
VN	verb, past participle
WP	Wh-pronoun
WRB	Wh-adverb

POS Abbreviation

POS	Abbreviation
Noun	N
Verb	V
Preposition	P
Adjective	A
Adverb	Adv
Conjunction	C
Pronoun	Pro

APPENDIX D. THE TOKENIZATION SPECIFICATIONS OF MA IN XML FRAGMENTS

```
<?xml version="1.0" encoding="utf-8"?>
<!--
 ~ Copyright (c) 2013. The Trustees of Columbia University in the City of New York.
 ~ The copyright owner has no objection to the reproduction of this work by anyone for
 ~ non-commercial use, but otherwise reserves all rights whatsoever.  For avoidance of
 ~ doubt, this work may not be reproduced, or modified, in whole or in part, for commercial
 ~ use without the prior written consent of the copyright owner.
 -->
<madamira_configuration xmlns="urn:edu.columbia.ccls.madamira.configuration:0.1">
  <preprocessing sentence_ids="false" separate_punct="true" input_encoding="UTF8"/>
  <overall_vars output_encoding="UTF8" dialect="MSA" output_analyses="TOP"
        morph_backoff="NONE"/>
  <requested_output>
    <req_variable name="PREPROCESSED" value="true" />
    <req_variable name="STEM" value="true" />
    <req_variable name="GLOSS" value="true" />
    <req_variable name="LEMMA" value="true" />
    <req_variable name="DIAC" value="true" />
    <req_variable name="ASP" value="true" />
    <req_variable name="CAS" value="true" />
    <req_variable name="ENC0" value="true" />
    <req_variable name="ENC1" value="false" />
    <req_variable name="ENC2" value="false" />
    <req_variable name="GEN" value="true" />
    <req_variable name="MOD" value="true" />
    <req_variable name="NUM" value="true" />
    <req_variable name="PER" value="true" />
    <req_variable name="POS" value="true" />
    <req_variable name="PRC0" value="true" />
    <req_variable name="PRC1" value="true" />
    <req_variable name="PRC2" value="true" />
    <req_variable name="PRC3" value="true" />
```

```
    <req_variable name="STT" value="true" />
    <req_variable name="VOX" value="true" /
    <req_variable name="BW" value="false" />
    <req_variable name="SOURCE" value="false" />
        <req_variable name="LENGTH" value="true" />
        <req_variable name="OFFSET" value="true" />
  </requested_output>
  <tokenization>
    <scheme alias="ATB" />
    <scheme alias="ATB4MT" />
    <scheme alias="MyD3">
      <!-- Same as D3 -->
      <scheme_override alias="MyD3"
                form_delimiter="\u00B7"
                include_non_arabic="true"
                mark_no_analysis="false"
                token_delimiter=" "
                tokenize_from_BW="false">
        <split_term_spec term="PRC3"/>
        <split_term_spec term="PRC2"/>
        <split_term_spec term="PART"/>
        <split_term_spec term="PRC0"/>
        <split_term_spec term="REST"/>
        <split_term_spec term="ENC0"/>
        <token_form_spec enclitic_mark="+"
                  proclitic_mark="+"
                  token_form_base="WORD"
                  transliteration="UTF8">
          <normalization type="ALEF"/>
          <normalization type="YAA"/>
          <normalization type="DIAC"/>
          <normalization type="LEFTPAREN"/>
          <normalization type="RIGHTPAREN"/>
        </token_form_spec>
      </scheme_override>
    </scheme>
  </tokenization>
</madamira_configuration>
```

APPENDIX E. EXAMPLES OF EXTRACTED POS PATTERNS

2 Grams Pattern	Frequency	3 Grams Pattern	Frequency	4 Grams Pattern	Frequency
IN NN	67708159	NN PRP$ NN	22266146	NN PRP$ NN DTNN	5060546
PUNC CC	56202948	DTNN PUNC CC	16304513	NN DTNN PUNC CC	4756362
IN PRP	41366327	IN NN DTNN	14269126	NN PRP$ NN PRP$	4444278
DTNN PUNC	39256456	NN NN DTNN	14171604	NN DTNN CC DTNN	4210429
DTNN CC	38925983	NN PRP$ DTNN	13987631	DTNN NN PRP$ NN	3934838
CC NN	37373483	IN NN NN	13419895	NN PRP$ NN NN	3676491
CC VBD	33334716	DTNN CC DTNN	13277922	DTNN CC DTNN CC	3419841
DTNN NN	31662332	NN NN PRP$	12863556	NN DTNN NN PRP$	3320738
DTNN DTJJ	29176791	PUNC CC VBD	12333425	NN NN PRP$ NN	3299981
DTNN IN	28480581	NN DTNN CC	12193994	DTNN PUNC CC VBD	3272982
VBP NN	26845794	DTNN IN NN	11593557	NN PRP$ PUNC CC	3241182
NN NNP	25660813	DTNN NN PRP$	11475177	NN DTNN IN NN	3058919
CC DTNN	23975751	NN DTNN PUNC	11043746	DTNN NN PRP$ DTNN	3006488
VBD NN	23234621	NN NN NN	10937729	IN NN NN DTNN	2830782
IN DTNN	23103478	NN DTNN DTJJ	10142851	DTNN IN NN DTNN	2791289
PRP$ NN	22831537	NN IN NN	9206645	PUNC CC VBD NN	2782322
NN IN	22704829	IN PRP VBP	8711751	NN DTNN CC NN	2772469
NN PRP$ NN	22266146	NN DTNN NN	8391728	DTNN CC DTNN PUNC	2705460
NNP NN	22220348	IN PRP NN	8195764	NN PRP$ CC NN	2646934
DTNN DTNN	22131052	NN PUNC CC	7733775	DTNN IN NN NN	2458084
NN PUNC	21575200	CC VBD NN	7661824	IN NN DTNN CC	2456009
NN CC	20050017	PUNC CC VBP	7444318	NN NN NN DTNN	2442132

2 Grams pattern	frequency	3 Grams pattern	frequency	4 Grams pattern	frequency
CC VBP	18857376	PUNC CC NN	7320440	NN PRP$ IN NN	2414761
VBP IN	18064816	NN PRP$ PUNC	7274713	NNP IN NN NNP	2413652
CC IN	17950267	NN DTNN IN	7152758	NN NN DTNN CC	2389006
PRP NN	17506592	NN DTNN DTNN	6924347	DTNN DTJJ PUNC CC	2385316
NN JJ	17216501	CC NN DTNN	6900405	PUNC CC VBD DTNN	2378600
PUNC NN	16900220	NN PRP$ CC	6898072	CC DTNN CC DTNN	2291899
CD CD	16852224	CC NN PRP$	6866364	DTNN PUNC CC VBP	2218724
VBP DTNN	16522310	DTNN CC NN	6820013	IN NN NN PRP$	2214410
DTNN PUNC CC	16304513	IN NN NNP	6798791	DTNN PUNC CC NN	2212689
NNP NNP	15963921	NN CC NN	6732523	IN NN DTNN PUNC	2191304
WP VBP	15830094	VBP NN PRP$	6668608	NN NN DTNN DTJJ	2184535
DTNN WP	14605239	NNP IN NN	6627166	IN NN NN NN	2183486
NNP DTNN	14504285	CD CD CD	6505943	NN IN NN DTNN	2141883
VBP PRP	14418345	CC NN NN	6423179	NN NN PRP$ DTNN	2117454
IN NN DTNN	14269126	VBP IN PRP	6098589	CC IN PRP VBP	2115962
NN NN DTNN	14171604	IN NN PRP$	6054524	DTNN PUNC CC IN	2109085
PRP$ DTNN	14123049	VBP IN NN	5909778	VBP NN PRP$ NN	2083282
NN PRP$ DTNN	13987631	VBD NN PRP$	5792764	NN NN DTNN PUNC	2077123
VBD DTNN	13899827	DTNN DTJJ PUNC	5645311	NN PRP$ DTNN CC	2065963
DTNN NNP	13849042	IN WP VBP	5640551	NN PRP$ DTNN DTJJ	2058331
NNP IN	13637967	CC VBD DTNN	5581847	DTNN DTJJ IN NN	2050542
VBD IN	13608850	CC IN PRP	5580951	DTNN CC NN DTNN	2007566
NNP PUNC	13491786	NN PRP$ NNP	5518467	IN PRP VBP NN	1964491
VBP VBP	13449635	NN PRP$ IN	5426477	CD CD CD CD	1958490
CD PUNC	13441110	VBD IN PRP	5256889	PUNC PUNC PUNC PUNC	1956649
IN NN NN	13419895	PRP$ NN DTNN	5109266	CC VBD NN PRP$	1954090
DTNN CC DTNN	13277922	DTNN IN DTNN	4902224	NN DTNN DTJJ PUNC	1933275
NN NN PRP$	12863556	VBP NN DTNN	4849487	NN NN NN PRP$	1881308
CD NN	12534688	CC DTNN CC	4712831	IN NN DTNN DTJJ	1875270

5 Grams Pattern	Frequency	6 Grams Pattern	Frequency
NN PRP$ CC NN PRP$	1480560	DTNN CC DTNN CC DTNN CC	655452
VBD NNP IN PRP CC	1348821	CC DTNN CC DTNN CC DTNN	481025
DTNN CC DTNN PUNC CC	1313136	NNP VBD NNP IN PRP CC	467528
DTNN NN PRP$ NN DTNN	1108906	NN DTNN CC DTNN CC DTNN	449107
NN DTNN NN PRP$ NN	1106830	NN PRP$ NN DTNN PUNC CC	434487
NNP IN PRP CC VBD	1106401	NN NNP VBD NNP IN PRP	432482
NN PRP$ DTNN PUNC CC	1106400	DTNN VBD NNP IN PRP CC	427348
NN PRP$ NN PRP$ NN	1056792	NNP IN NN NNP IN NN	405891
IN NN DTNN PUNC CC	996836	NN DTNN CC DTNN PUNC CC	401749
NN DTNN PUNC CC VBD	988905	DTNN CC DTNN CC DTNN PUNC	387440
NN PRP$ NN DTNN PUNC	955389	NN PRP$ CC NN PRP$ CC	382054
NN DTNN CC DTNN CC	948796	NN PRP$ CC NN PRP$ PUNC	373608
NN DTNN NN PRP$ DTNN	946560	CD CD CD CD CD CD	373537
NN NN DTNN PUNC CC	914671	NNP IN PRP CC VBD PUNC	364542
NN DTNN CC NN DTNN	905248	IN NN NNP IN NN NNP	354018
NN PRP$ NN DTNN CC	890883	CC NN PRP$ CC NN PRP$	344254
NN PRP$ NN NN DTNN	860818	NN DTNN NN PRP$ NN DTNN	326775
NN DTNN DTJJ PUNC CC	840369	NN PRP$ NN DTNN CC DTNN	292940
NN DTNN CC DTNN PUNC	835322	NN NN PRP$ CC NN PRP$	286418
CC DTNN CC DTNN CC	813649	CD PUNC CD PUNC CD CD	264406
PUNC CC IN PRP VBP	804411	CC DTNN CC DTNN PUNC CC	253302
CD CD CD CD CD	790831	NN PRP$ NN PRP$ NN DTNN	252125
NN NN DTNN CC DTNN	788682	DTNN NN PRP$ DTNN PUNC CC	251668
IN NN DTNN CC DTNN	780480	VBD NNP IN PRP CC NN	250341
NN NN PRP$ NN DTNN	767163	NN PRP$ NN PRP$ PUNC CC	249809
DTNN PUNC CC VBD NN	766056	IN NN PRP$ CC NN PRP$	249755
NN PRP$ NN PRP$ DTNN	743120	NN DTNN DTJJ IN NN DTNN	244378
NN DTNN DTJJ IN NN	741815	CD CD PUNC CD PUNC CD	244106
NN DTNN IN NN DTNN	740985	NN DTNN PUNC CC VBD NN	240409
NN NN PRP$ NN PRP$	731705	NN NN DTNN NN PRP$ NN	238179
NN DTNN PUNC CC NN	731430	NN PRP$ DTNN DTJJ PUNC CC	237990
NN NN DTNN NN PRP$	731144	NN PRP$ DTNN PUNC CC VBD	236048
NN PRP$ NN DTNN DTJJ	721134	DTNN PUNC CC IN PRP VBP	229446
CD PUNC CD PUNC CD	719630	NN DTNN DTJJ NN PRP$ NN	225482
NN DTNN DTJJ NN PRP$	709362	NN DTNN CC DTNN NN PRP$	224886
IN NN NNP IN NN	704847	NN DTNN CC NN DTNN CC	224831

5 Grams Pattern	Frequency	6 Grams Pattern	Frequency
NN DTNN IN NN NN	679018	NN DTNN PUNC CC VBD DTNN	217582
DTNN NN PRP$ NN NN	670080	NN NN DTNN NN PRP$ DTNN	215970
DTNN DTJJ IN NN DTNN	668610	PRP$ CC NN PRP$ CC NN	215116
NN NN DTNN CC NN	656066	IN NN DTNN PUNC CC VBD	213547
NN PRP$ DTNN CC DTNN	647435	IN NN DTNN NN PRP$ NN	212417
CC NN PRP$ PUNC CC	646861	PUNC CC VBD NN PRP$ NN	211985
NN NN PRP$ PUNC CC	642366	DTNN IN NN DTNN PUNC CC	209823
IN NN DTNN NN PRP$	638032	NN DTNN PUNC CC NN DTNN	208545
NN PRP$ NN NN PRP$	632071	NN NN DTNN CC NN DTNN	208162
NN DTNN PUNC CC VBP	631682	NN NN DTNN PUNC CC VBD	203270
PUNC CC VBD NN PRP$	631387	PRP$ CC NN PRP$ PUNC CC	201471
NN NN DTNN IN NN	626423	DTNN CC DTNN NN PRP$ NN	201397
IN NN DTNN IN NN	624647	DTNN NN PRP$ NN DTNN CC	200001
NN DTNN PUNC CC IN	606646	NN PRP$ NN DTNN CC NN	199152
NN DTNN CC NN PRP$	606143	NN DTNN DTJJ PUNC CC VBD	197333
NN PRP$ IN NN DTNN	602094	DTNN CC DTNN CC NN DTNN	195936
PUNC CC IN WP VBP	601216	NN PRP$ NN PRP$ NN NN	194048
DTNN DTJJ NN PRP$ NN	590039	NN PRP$ NN DTNN IN NN	192078
NNP VBD NNP IN PRP	586341	IN NN NN DTNN PUNC CC	190589
CC VBD NN PRP$ NN	584547	NN DTNN NN PRP$ NN NN	190440
DTNN NN PRP$ NN PRP$	580616	NN DTNN DTJJ NN PRP$ DTNN	190275
DTNN CC DTNN NN PRP$	579248	DTNN CC DTNN PUNC CC VBP	187973
NN PRP$ NN NN NN	573343	NN NN PRP$ NN PRP$ NN	187144
NN PRP$ NN PRP$ PUNC	568261	NN DTNN CC NN DTNN PUNC	187052
NN NN PRP$ CC NN	568231	NN NN PRP$ DTNN PUNC CC	186135
DTNN IN NN NN DTNN	563293	PUNC CC IN PRP VBP NN	184727
PUNC CC VBD NN DTNN	556739	NN PRP$ CC NN NN PRP$	183621
IN NN DTNN CC NN	556141	DTNN DTJJ NN PRP$ NN DTNN	182388
PUNC CC VBD IN PRP	550647	CD CD PUNC CD CD CD	180956
NN NNP IN NN NNP	544261	NN NN DTNN DTJJ PUNC CC	180486
NN PRP$ DTNN DTJJ PUNC	542100	DTNN CC DTNN NN PRP$ DTNN	180459
DTNN PUNC CC NN DTNN	538727	DTNN CC DTNN PUNC CC NN	179879

APPENDIX F. EXAMPLES OF EXTRACTED INSTANCES OF AMWE

DTNN-DTJJ	Frequency	NN-DTNN	Frequency
الامم المتحدة	2050	حقوق الانسان	42236
المجتمع الدولي	851	رئيس الوزراء	21145
الاجهزة الامنية	784	كرة القدم	18326
العام الماضي	720	مجلس الوزراء	16905
القضية الفلسطينية	607	رئيس الجمهورية	13938
الحكومة العراقية	607	يوم القيامة	13789
الشعب السوري	606	مجلس النواب	13076
الوحدة الوطنية	577	سبيل المثال	12615
الاتحاد الاوروبي	574	مجلس الامن	12409
القوى السياسية	566	اهل السنة	11245
العملية السياسية	540	نفس الوقت	10973
المجلس الاعلى	500	اهل البيت	10690
المادة السابقة	500	يوم الجمعة	10190
النظام السوري	498	وزير الخارجية	9187
الشريعة الاسلامية	487	اهل العلم	8797
المجلس الوطني	467	خلال الفترة	8583
القانون الدولي	454	دول العالم	7881
التعليم العالي	453	ن الخطاب	7833
المجلس العسكري	426	غض النظر	7805
الكيان الصهيوني	424	زارة الداخلية	7644
النظام السابق	418	رئيس المجلس	7335
الفقرة السابقة	409	بين الناس	7125
البحث العلمي	375	يوم السبت	6934
الاحزاب السياسية	366	مجلس الشعب	6813
المحكمة الدولية	362	يوم الاثنين	6665
العملية الانتخابية	359	يوم الخميس	6652
الكتل السياسية	329	يوم الاحد	6596
العام الحالي	325	اطلاق النار	6353
الرئيس السوري	321	يوم الاربعاء	6187
العالم الاسلامي	316	انحاء العالم	6105
البنية التحتية	313	عض الاحيان	6067
المحكمة الجزائية	310	مع العلم	5986
المواد الغذائية	309	يوم الثلاثاء	5713

DTNN-DTJJ	Frequency	NN-DTNN	Frequency
الحوار الوطني	294	مجلس الادارة	5567
الرئيس السابق	286	رئيس الحكومة	5489
الخدمة المدنية	286	زارة التربية	5419
الدولة الفلسطينية	271	صباح اليوم	5376
المرحلة المقبلة	270	ميدان التحرير	5250
المدينة المنورة	265	رجال الاعمال	5137
الامن الدولي	265	ارض الواقع	5078
الفترة الماضية	259	راس المال	5048
الثانوية العامة	259	امن الدولة	4803
الاراضي الفلسطينية	258	جلالة الملك	4741
الفترة المقبلة	251	رئيس اللجنة	4631
الدين الاسلامي	251	اعادة النظر	4591
القائمة العراقية	247	منظمة التحرير	4571
الشهر الجاري	247	خلال العام	4564
الادارة الامريكية	246	عض الناس	4532
العلي القدير	242	مجلس الشورى	4459
العمل السياسي	235	زارة الصحة	4453
الهيئة العامة	233	اس العالم	4443
الوقت الحالي	230	وزير الدفاع	4437
الطاقة الكهربائية	229	شيخ الاسلام	4432
الحكومة السورية	228	وكالة الانباء	4429
الشرعية الدولية	227	جامعة الدول	4417
المؤتمر الوطني	226	رجال الدين	4375
العام المقبل	224	جماعة الاخوان	4196
الجمهورية الاسلامية	224	حرية التعبير	4144
البنك المركزي	224	حزب البعث	4093
المصالحة الوطنية	223	منطقة الشرق	3996
الحكومة الاسرائيلية	223	ابناء الشعب	3933
العام الجاري	222	ضغط الدم	3930
التعاون الخليجي	222	اهل الكتاب	3880
الشهر الماضي	218	عض الدول	3863
الامن القومي	218	فضيلة الشيخ	3832
المبادرة الخليجية	215	اعضاء المجلس	3831
الحياة السياسية	214	حول العالم	3826

APPENDIX G. EXAMPLES OF TEST DATA AND ANNOTATIONS TEST

Nu.	candidates		anno.1	anno.2	comment	dataset
1	تتأتي في اطار	tataʾtī fī aṭār	1	1		V3
2	يصب في مصلحة	yaṣb fī maṣlḥa	1	1		V3
3	حسبنا الله عليهم	ḥasbnā ʾallāh ʿalīhm	1	0		N4
4	يتواكب مع	yatwākb maʿ	1	1		V2
5	تتراوح بين	tatrāwḥ bayn	1	1		V2
6	بمعنى اخر	bamʿnā aẖr	1	1		P3
7	مصادر مطلعة	maṣādr maṭlʿa	1	1		N2
8	تلبية احتياجات	talbya aḥtyājāt	1	1		N2
9	دون حسيب او رقيب	dawn ḥasīb aw raqīb	1	1		P3
10	ارض الله واسعة	arḍ ʾallāh wasʿa	1	1		N3
11	توخي الحيطة و الحذر	tawẖī alḥayṭa wa alḥaḏr	1	1		N4
12	صلاح الدنيا و الدين	ṣalāḥ addanyā wa addayn	0	1		N4
13	على سبيل المثال	ʿalā sabīl almaṯāl	1	1		P3
14	على ارض الواقع	ʿalā arḍ alwāqʿ	1	1		P3
15	على مدار الساعة	ʿalā madār assāʿa	1	1		P3
16	بالنسبة	bālnsba	1	1		P2
17	بالتالي	bāltālī	1	1		P2
18	بالاضافة	bālāḍāfa	1	1		P2
19	بواسطة	bawāsṭa	0	1		P2
20	ذات اليمين و ذات الشمال	ḏāt alyamīn wa ḏāt aššamāl	1	0		N5
21	غسل الاموال	ġasl alāmwāl	1	1		N2
22	حامل المسك و نافخ الكير	ḥāml almask wa nāfẖ alkayr	1	1		N5
23	تبييض الاموال	tabyīḍ alāmwāl	1	1		N2
24	بمنزلة الراس من الجسد	bamnzla arrās man aljasd	1	1		P5
25	شحذ الهمم	šaḥḏ alhamm	1	1		N2
26	طول البلاد وعرضها	ṭawl albalād waʿrḍhā	1	1		N5
27	صلى الله عليه وسلم	ṣalā ʾallāh ʿalīh waslm	1	0		V5
28	من أي وقت مضى	man ʾay waqt maḍā	1	0		P4
29	على العكس من ذلك	ʿalā alʿax man ḏalk	1	1		P4
30	كل من تسول له نفسه	kal man tasūl lah nafsh	1	1		N6

APPENDIX H. XML FRAGMENT FOR THE AMWE, *FĪ ʾAMAS ALḤĀJAT*, في أمس الحاجة.

```
<LexicalEntry mwePattern="PreAdvNo">
        <feat att="partOfSpeech" val="preposition"/>
        <Lemma>
          <feat att="writtenForm" val=" في أمس الحاجة"/>
            </Lemma>
        <ListOfComponents>
          <Component entry="A1"/>
          <Component entry="A2"/>
          <Component entry="A3"/>
        </ListOfComponents>
    </LexicalEntry>
    <LexicalEntry id="A1" morphologicalPatterns="AsTable">
        <feat att="partOfSpeech" val="prepostion"/>
        <Lemma>
          <feat att="writtenForm" val="في"/>
        </Lemma>
    </LexicalEntry>
    <LexicalEntry id="A2" morphologicalPatterns="AsTable">
        <feat att="partOfSpeech" val="verb"/>
        <Lemma>
          <feat att="writtenForm" val="آمس"/>
        </Lemma>
    </LexicalEntry>
    <LexicalEntry id="A3" morphologicalPatterns="AsTable">
         <feat att="partOfSpeech" val="noun"/>
        <Lemma>
          <feat att="writtenForm" val="الحاجة"/>
        </Lemma>
    </LexicalEntry>
<MWEPattern id="NdeFixedN">
      <MWENode>
```

```
        <feat att="syntacticConstituent" val="NP"/>
        <MWELex>
          <feat att="rank" val="1"/>
          <feat att="graphicalSeparator" val="space"/>
          <feat att="structureHead" val="yes"/>
        </MWELex>
        <MWELex>
          <feat att="rank" val="2"/>
          <feat att="graphicalSeparator" val="space"/>
        </MWELex>
        <MWELex>
          <feat att="rank" val="3"/>
          <feat att="graphicalSeparator" val="space"/>
          <feat att="grammaticalNumber" val="singular"/>
        </MWELex>
    </MWENode>
</MWEPattern>
<LinguisticFeatures>
    <OrthographicFeatures>
        <feat att="Id" val="mwe1"/>
        <feat att="Comment" val=" "/>
        <feat att="DIN31635InPlainEnglish" val="fī ʾamasi alḥājat "/>
        <feat att="Normalised Form" val="في امس الحاجة"/>
        <feat att="Different Spelling Form" val=" "/>
    </OrthographicFeatures>
```