Towards Comprehensive Computational Representations of Arabic Multiword Expressions

Ayman Alghamdi[1] and Eric Atwell[2]

[1] Umm Al-Qura University, Makkah, KSA
[2] University of Leeds, Leeds, UK
scaaa@leeds.ac.uk
e.s.atwell@leeds.ac.uk

Abstract. A successful computational treatment of multiword expressions (MWEs) in natural languages leads to a robust NLP system which considers the long-standing problem of language ambiguity caused primarily by this complex linguistic phenomenon. The first step in addressing this challenge is building an extensive reliable MWEs language resource LR with comprehensive computational representations across all linguistic levels. This forms the cornerstone in understanding the heterogeneous linguistic behaviour of MWEs in their various manifestations. This paper presents a detailed framework for computational representations of Arabic MWEs (ArMWEs) across all linguistic levels based on the state-of-the-art lexical mark-up framework (LMF) with the necessary modifications to suit the distinctive properties of Modern Standard Arabic (MSA). This work forms part of a larger project that aims to develop a comprehensive computational lexicon of ArMWEs for NLP and language pedagogy LP (JOMAL project).

Keywords: Multi-Word Expressions, Language Resource, Computational Representations, Annotation.

1 Introduction

Multi-Word Expressions MWEs are a heterogeneous phenomenon in human languages which pose different types of serious challenges particularly in the fields of Natural Language Processing NLP and in language pedagogy LP. This is because MWEs are considered as one of the key factors that contribute to the long-standing language ambiguity problems which is one of the most crucial setbacks that most NLP tasks face. Sag, Baldwin, Bond, Copestake, and Flickinger (2002, p. p.15) emphasise that ‘like the issue of disambiguation, MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed’. However, in recent years much research has been conducted in this area which aims to scientifically study this phenomenon to discover new methods and approaches that aim to determine the best computational practice in MWEs processing based on state-of-the-art techniques and tools developed in NLP, machine learning and artificial intelligence research. These efforts cover a wide range of topics related to MWEs which includes but is not limited to extraction models, computational representations, linguistic analysis and classifications. The first step and the foundation stone of these studies is the availability of reliable representative open source MWE LRs which pave the way for interested researchers to experiment and

analyse various types of these lexical units in order to find out the best computational treatment and ultimately improve various NLP applications output.
Moreover, these LRs can be embedded in the implementations of various NLP and ML tools that take MWEs knowledge into account and which assist considerably in the task of tackling language ambiguity related problems. However, while many well developed MWEs lexicons are freely available for English and other modern European languages, Arabic is still suffering from lack of computational comprehensive MWE LRs. Most of the existing Arabic MWEs lexicons are either very limited in terms of their scale or annotations features or they are not freely available as an open source project which makes it difficult for most researchers to benefit from them. This project aims to remedy this deficiency by constructing a large scale Arabic MWEs lexicon with detailed computational representations at different linguistic levels based on state-of-the-art extraction methods and international standards for MWE computational representations. The current paper reports part of this project which focuses on describing a framework for computational representations and annotations across various linguistic levels.

2 Related work

Several projects have attempted to create an electronic database for different types of MWEs developed for various purposes; for instance, the SIGLEX-MWE website lists more than 22 MWE LRs in different languages which are open source projects that are available for free download[1]. In their on-going project which is led by a multidisciplinary scientific network devoted to European MWEs (Sag et al., 2002; Savary et al., 2015) provide a summary about their latest research results and activities related to the computational treatments of MWEs. However, as part of their project, Losnegaard et al. (2016)) and Rosén et al. (2016) conducted an intensive survey of all the available MWEs LRs based on the use of an online questionnaire which was designed to obtain detailed information about all the existing electronic MWEs resources. The preliminary results of the survey are publicly accessible as an online updated spreadsheet[2]. Based on the main classifications of the study questionnaire[3] with minor modifications, the MWEs lexicons were grouped into five categories as can be seen in table 1.

Table 1. The Main categories of MWE LRs

MWEs lexicon	LRs Nu	Percentage
Treebank with MWE annotations	12	11%
MWE lexicons	48	45%
Monolingual list of MWEs	13	12%
Multilingual resources	15	14%
Others (for all the LRs out of the previous categories)	19	18%

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

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

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

These MWEs lexicons represent a variety of LRs types in relation to their domain, phrase length, linguistic annotation, size and degree of accessibility; however the result related to Arabic MWEs lexicons shows only four MWE LRs developed by (Al-Sabbagh, Girju, & Diesner, 2014; Arts, 2014; Cardey, Chan, & Greenfield, 2006; Steinberger, Pouliquen, Kabadjov, & Van der Goot, 2013). Unfortunately, they are not publicly accessible through the web which make it difficult for us to find out more details about their scale or what type of linguistic annotation is associated with them.

Another Arabic MWEs list was developed by Attia (2006) in the process of creating an Arabic version of the Xerox Linguistic platform which was initially developed by Butt (1999) and Dipper (2003) for writing languages grammar rules and performing various linguistic analyses based on the Lexical Functional Grammar theory (Wanner, 1996). In the process of building the MWE Transducer, Attia managed to extract a list of 2826 Arabic MWEs items which were then classified into four main categories based on the classifications of MWEs presented by (Sag et al., 2002), as can be seen in Table 2 along with examples:

Table 2. Arabic MWEs classifications with examples.

Classifications	ArMWEs	Translation
Compositional expressions	*ġalāf alkatāb*[4], غلاف الكتاب	Book cover
Non-compositional expressions	*raj‘ bak̮fī ḥanīn*,رجع بخفي حنين	Kick the bucket
Fixed Expressions	*ʾišāra almarūr* ,إشارة المرور	Traffic light/lights
Semi-fixed Expressions	*ṣabāḥ alk̮ayr*, صباح الخير/الخيرات	Good morning
Flexible Expressions	*tak̮fīḍ sar‘a almarkba*, تخفيض سرعة المركبة	Slow the car down

In this project, several types of MWEs are excluded from the extracted list including compound nouns, verbal and prepositional phrases. In addition, there was no intention to create lexical representations for MWEs listed because the sole aim was to improve the linguistic analyser system by accommodating several types of MWEs. It is worth mentioning here that the experiment's findings in this research emphasise the major positive role of accommodating MWEs LR in the final system output which was less ambiguous with a higher degree of precision which again highlights the importance of creating a large scale Arabic MWE LR which will help in the improvement of many Arabic NLP tasks.

Bounhas and Slimani (2009), implemented another hybrid model for extracting compound nouns and also proposed new algorithms to reduce morphological and syntactic ambiguities during the MWE extraction process. Their model was constituted of three phases starting from the morphological analysis, followed by the sequence identifier and syntactic parser. The final result was filtered based on statistical information. The extracted items were classified into six categories according to different types of Arabic compound nouns as shown in table 3.

[4] The German standard DIN 31636 is used for rendering Romanized Arabic

Table 3. Arabic compound nouns classifications with examples

Compound noun classifications	Examples	Examples
Annexation compound noun	سيارة رجل غني	The car of a rich man
Adjective compound noun	بيت كبير	Big house
Substitution compound noun	هذه السيارة	This car
Prepositional compound noun	نوع من الحلوى	A kind of sweet
Conjunctive compound noun	القط والفأر	The cat and the mouse
Compound nouns linked by composite relations	الاستمرار لحوالي سنة	To persist for about one year

To evaluate their model, the final list of MWEs was compared to a previously well-developed MWEs list and the result shows an improvement in the extracting accuracy in comparison with previous experiments applied on Arabic MWEs extractors in the same domain.

A more recent study carried out by (Hawwari, Bar, & Diab, 2012) aimed to build a list of MWEs collected manually from existing written Arabic MWEs dictionaries to automatically annotate an Arabic corpus using a pattern-matching algorithm to help in the automatic statistical identification of MWEs in running text. Their final list was categorised into five groups based on syntactic constructions as can be seen in Figure 1.

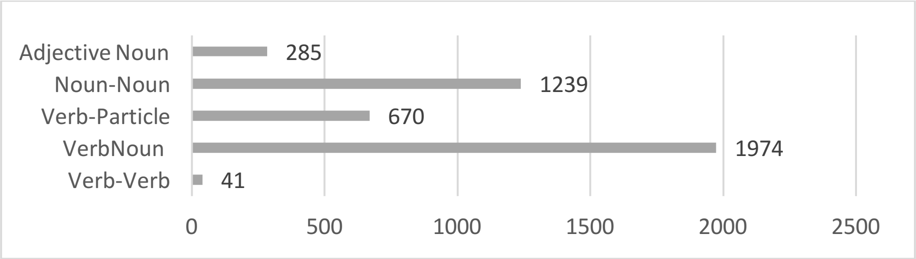

Fig. 1. Syntactic constructions of Arabic MWES in (Hawwari et al., 2012).

In their following study (Hawwari, Attia, & Diab, 2014) presented a framework for the classification and annotation of Egyptian Arabic MWEs. Their focus was on representing different types of MWEs in the Egyptian dialect. It is worth considering here that several of the annotation features suggested in this study are applicable with several modifications to MWEs in MSA which are this study’s main concern. The classifications and annotation cover different linguistic levels such as morphological, syntactic, semantic and pragmatic features of MWEs. The developed framework builds on previous research applied to other languages e.g. (Calzolari et al., 2002; Tanabe, Takahashi, & Shudo, 2014). Another study by Al-Sabbagh et al. (2014) aims to build Arabic modal multiword expressions to accelerate the automatic extraction process and represents variation patterns of modal MWEs in Arabic. Based on (Palmer, Gildea, &

Kingsbury, 2005)'s cross-lingual taxonomy of modality senses, they classify the extracted MWEs items into 7 categories as follows: (un)certainty, evidentiality, obligation, permission, commitment, ability and volition. Table 4 shows examples of MWEs classifications. Although, the author stated that the final LR is available for free download, we could not find a copy or working link for this MWE lexicon.

Table 4. Examples of AM-MWEs from Al-Sabbagh et al. (2014)

AM-MWEs	English translations
نويت	I intend
يمكنني	I can
أعتقد	I think
هناك احتمال بأن	It is possible

Bar, Diab, and Hawwari (2014) developed a relatively small manually annotated list of MWEs as a gold standard list in the process of tackling the problem of automatic extraction and classifications of MWEs. They implemented deterministic pattern-matching algorithms in the detection process of various types of continuous and non-continuous MWEs; they found that the use of only shallow annotated data results in major improvements in the automatic boundary detection on the token level of MWEs.
Overall, in the developments of the current lexicon we build on all the previous attempts in the extractions and the lexical representations of ArMWEs, aiming to reach an innovative large scale Arabic MWE lexicon with comprehensive computational representations at various linguistics levels.

3 Properties of MWE Computational representations

Based on the main project objectives, the annotation scheme had to be easy to integrate in different types of NLP systems, following the state-of-the-art standards in lexical mark-up research. In addition, the adopted scheme is not restricted to any particular grammatical framework because of the reusability purposes 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 the current representations is the flexibility which cuts across all types of ArMWEs and also covers discontinuous as well as contiguous phrases; it is also targeted to be human readable and equally adopted for NLP systems to accommodate different end users' needs. However, most of the previous studies on ArMWEs annotation schemes have prioritised certain types of expressions or language genres to the exclusion of others, so they are not appropriate for representing multiple kinds of ArMWEs in our lexicon, which should allow for permutations across various linguistic levels. The computational ArMWEs representations are encoded in Extensible Mark-up Language XML because it is the most flexible and also the most used method in the representations of computational LR. The final version will be converted into HTML pages for the purpose of publishing its content on the Internet.

In our project we also benefit from the international standard lexical mark-up framework LMF which was the result of 60 experts' contributions who worked for more than five years to develop lexical representations and standards for different types of computational LRs (Francopoulo, 2013; Francopoulo & 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 & George, 2008). It is important to note that adopting standardisation when building computational LR can be very beneficial specially in NLP oriented applications as Francopoulo (2013, p. 3) showed:

> *'The significance of standardization was thus recognized, in that it would open up the application field, allow an expansion of activities, sharing of expensive resources, re-use of components and rapid construction of integrated, robust, multilingual language processing environments for end-user'*

Furthermore, the developed representations system gives special 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. p.4) states that 'For developers of general-purpose corpus resources, the aim may be to enrich the text with linguistic analyses to maximize the potential for corpus reuse in a wide range of applications'. In the following, a brief description of the type of users targeted in the JOMAL project is presented. This is followed by a detailed illustration of the adopted ArMWEs classifications and representations across different linguistic levels.

4 JOMAL Computational Representations

As mentioned previously, in the design of our lexicon annotation and classifications, this project takes into account the LMF core package and MWE patterns extension with the necessary deviations to facilitate the JOMAL reusability and connectivity to other LRs and various NLP systems and applications. This section describes the computational representations and the labels adopted for each MWEs class and propriety property with examples from Arabic corpora.

Throughout, we have made as much use of automated procedures as possible to reduce the time and effort of 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.

4.1 Basic lexicon information

This class is mainly adopted from the MWE extension in LMF framework, and it expresses the main information about the JOMAL which can be useful for the 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 which involves written form, related form and lexicographic type. Other classes aim to represent the details of MWE components in their various linguistic manifestations.

Table 5. Basic lexicon information representations in JOMAL.

Class Name	Subclasses and attributes
Lexical Resource	
Global Information	Label Comment Author Language Coding Script Coding

As can be seen in Table 5, the ID attribute which can be seen in most annotation classes was created to facilitate the linkage between shared annotation classes; thus they 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 2 shows an example of the XML fragment of the Global Information class:

```
<GlobalInformation>
     <feat att="label" val="Arabic Formulaic Sequences 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>
```

Fig.2. An example of lexicon information annotated in XML

4.2 Linguistic representations

The linguistic annotation classes are the core package of the JOMAL model which aims to provide a detailed linguistic description of each ArMWE in our lexicon. The annotations are classified into six main layers, each one is dedicated for linguistic levels starting from the shallow orthographic form of the lexical entry to the deep semantic and pragmatics features of MWE. The following subsections present a brief explanation of these linguistic annotations.

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 6.

Table 6. 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 class is the core class of each lexical entry and it is associated with all the annotation features. It also has several attributes related to written and related forms of MWE. For instance, the lexicographic types of the expressions represented by several labels as can be seen in Table 7 with examples from the lexicon.

Table 7. Examples of lexicographic types labels in JOMAL

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	اضرب الحديد وهو حامي	*Hit the iron while it's hot*

The MWEs pattern instance is a shared resource which provides information about different lexical combination phenomena. This class is associated and explained by the list of components instance that contain all the expression constituent words. The node classes' aim is to represent the structure 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 constituents of the English phrase to take off is Verb_ Preposition or VP; an Arabic equivalent example can be seen in the phrase, أخذ عن which is also classified as VP structure. In Table 8, examples of syntactic constituents found in JOMLA are listed.

The pattern type represents the degree of phrase morphological, lexical and grammatical flexibility by using a scale of three levels as illustrated in Table 9.

Table 8. Examples of syntactic constituents’ classifications in JOMLA

Label	**Example**
Noun_Noun	تكميم الأفواه, *takmīm al’afwāh*
Verb_Noun_Preposition_Noun	تجمد الدم في عروقه, *tajmd addam fī ‘arūqh*
Noun_Adjective	اليد المغلولة, *alyad almaġlūla*
Noun_Adverb	الأيام بيننا, *al’ayām baynnā*
Noun_Preposition	التغطية على, *attaġṭya ‘alā*
Preposition_Noun_Preposition	من أجل أن, *man ’ajl ’an*
Noun_Preposition_Noun	النوم في العسل, *annawm fī al‘asl*

Table 9. Pattern types classifications with Arabic examples

Flexibility degree	**Example**
Fixed MWE	رجع بخفي حنين, *raj‘ bakfī ḥanīn*
Semi-fixed MWE	أثلج/أثلجت صدره/صدرها, *’aṯlj/a’ṯljt ṣadrh/ṣdrhā*
Flexible MWE	أثقلته/أثقله/أنهكته الأعباء/الحمل/المسؤوليات، *’aṯqlth/a’ṯqlh/a’nhkth al’a‘bā’/alḥml/almsu’ūlyāt*

The MWE lex class used to provide a reference to each lexical component in the list of components instance. 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 one for other MWEs that have to be used with the same lexical items or what we called fixed MWEs. The Structure Head represents the first POS tag for the phrases and the rank attribute shows the components order and also any possible alternative orders. This feature is important particularly for Arabic because it has a high degree of flexibility in the order of sentence words. For instance, the MWE أقبلت عليه الدنيا has six components order possibilities shown in Table 10.

Table 10. An example shows the components order flexibility in ArMWEs

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

Orthographic representations

As described in Table 11, 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 they enable them to extract the LR in various formats according to the targeted NLP or ML tasks.

which can be represented in various forms based on its orthographic features, as in Table 12.

Table 11. The linguistic annotation layers of JOMAL

Class Name	**Subclasses and attributes**
Orthographic Features	Id Comment DIN31635RenderingInPlainEnglish Normalised Form Different Spelling Form
Phonological Features	Id Comment Diacritization 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
Pragmatics Features	Id Comment Usage Type User Type

Table 12. An example of orthographic features of MWE أعياه الأمر , *exhaust*

Orthographic Features	**Expression example**
DIN31635RenderingInPlainEnglish	*ʾaʿyāh alʾamr*
Normalised Form	اعياه الامر
Different Spelling Form	أعياه الأمر

To see an example of the previous annotation in XML, Appendix 1 illustrates the XML fragment which represents the ArMWE في أمس الحاجة,*fī ʾams alḥāja, in urgent need*.

Phonological representations

At the Phonological layer of annotation, we provide a complete diacritization of each phrase which is an essential feature in Arabic phonology to express the most common pronunciation form of ArMWEs in MSA. This representation is also particularly important because of the absence of short vowel symbols in Arabic script, which also play a prime role at the syntactic and semantic analysis levels of the lexical units. Other attributes are devoted to represent other phonological variants when available and also a representation of the expression in IPA phonetic script.

Morphosyntactic representations

For the Morphosyntactic representations we use a modified version of LMF morphological patterns extension 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 which result in different variations for each word which we aim to represent in JOMAL lexicon. With regard to the POS feature, expressions components are classified into five categories according to their POS tag. Table 13 shows the adopted morphological tag set with MWE examples of the headword POS.

Table 13. Examples of the POS tags used in the morphosyntactic representations

POS tag	Example
Noun	البرج العاجي *albarj alʿājī*
Verb	التزم الصمت *attazm aṣṣamt*
Adjective	جنون العظمة *janūn alʿaḓma*
Adverb	بين الحياة والموت *bayn alḥayā walmawt*
Preposition	على قدم المساواة *ʿalā qadm almasāwā*
Interjection	يا غالب يا مغلوب *yā ġālb 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 MWEs representations because of the derivational and inflectional nature of Arabic morphology which means that words in Arabic are derived from specific roots, and usually the inflected words that share the same root belong to a common semantic field. Thus this feature helps to easily classify all the words that belong to the same root into semantically similar groups based on the common morphological root. Table 14 shows an example of an Arabic root with its morphological patterns and inflection forms.

Table 14. Examples of Morphological patterns and meanings of the word سمع

Morphological patterns	Meaning
سمع	Listen(Past tense verb)
يسمع	Listening (Present tense verb)
اسمع	Listen (Imperative tense verb)
مسموع	Heard
سماعة	Speaker
سامع	Listener (Singular)
سامعون	Listeners (Plural for male)
سامعات	Listeners (Plural for female)

The grammatical features class is targeted to represent four main properties, including number, gender, tense for verbs and person. Consequently all these features involve several values which are represented in detail in the grammatical properties of each MWE component. Table 15 provides examples of these linguistic features in Arabic.

Table 15. Examples of Grammatical features annotation

Grammatical features	Values
Number	Signal, plural
Gender	Male, female, things
Tense	Past, present, imperative
Person	Third person

Semantic representations

This level of annotation constitutes four main classes created for representing the sematic information of MWEs. The ‘Sense Set’ class represents the meaning variants of MWEs in different contexts associated with a corpus example that reflects the real use of the phrase. The ‘Semantic Fields’ class aims to group the phrases into several categories based on the main semantic fields. The idiomaticity degree feature is targeted to classify the MWEs into three categories based on the ambiguity levels of the phrase as follows: full opaque, semi opaque and compositional MWEs. Full opaque MWEs involve expressions that have no semantic relation between the general meaning of the phrase as a whole and its component parts, such as, على كف عفريت، على قدم وساق، طالت أظافره، *ʿalā kaf ʿafrīt, ʿalā qadm wasāq, ṭālt ʾaḓāfrh.* Semantic Relations is a class representing the oriented relationship between Synset instances, where three types of relations are included: synonymy, antonymy, and polysemy.

Pragmatics 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 a MWEs’ pragmatic behaviour.

4.3 Pedagogical representations and other features

The aim of these representations is to make the most of JOMAL in any language pedagogy related applications. Thus this class provides valuable information in this regard which includes frequency attributes which show the popularity degree of the phrase. In addition, the source label presents information about the LRs where phrases were extracted from. 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 written or speech corpus. As listed in Table 16 the last class of our representations model was created to include all the information that are beneficial for the LR end-users and cannot belong to any of the previous described annotation classes; for instance the status of annotation compilation for each lexical entry and also the MWE Equivalent in Arabic dialects or the translation of MWE in other languages.

Table 16. 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

5 Conclusion and future work

In this paper, we present a detailed description of the lexical representations model that we applied in the development of a comprehensive ArMWEs lexicon for NLP and LP. In our model, we build on previous attempts and standards in the computational lexical representations of MWEs; moreover, we add several innovative annotation features that enhance the usefulness and the usability of JOMAL in various practical applications in NLP and LP. This work is a crucial and essential step towards more advanced and comprehensive research in the computational treatment of ArMWEs. This paper extends our earlier work on ArMWEs reported in (Alghamdi & Atwell, 2016, 2017). Future work will focus on building various tools and applications based on the developed lexicon to make the most out of it.

References

Al-Sabbagh, R., Girju, R., & Diesner, J. (2014). Unsupervised Construction of a Lexicon and a Repository of Variation Patterns for Arabic Modal Multiword Expressions. *EACL 2014*, 114.

Alghamdi, A., & Atwell, E. (2016). *An empirical study of Arabic formulaic sequence extraction methods*. Paper presented at the The10th International Conference on Language Resources and Evaluation, Portorož, Slovenia.

Alghamdi, A., & Atwell, E. (2017). Constructing a corpus-informed Listing of Arabic formulaic sequences ArFSs for language pedagogy and technology *Under review paper submitted to International Journal of Corpus Linguistics*.

Arts, T. (2014). *Oxford Arabic Dictionary: Arabic-English, English-Arabic*: Oxford University Press.

Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar (Vol. 4139, pp. 87-98). BERLIN: SPRINGER-VERLAG BERLIN.

Atwell, E. (2008). Development of tag sets for part-of-speech tagging. In A. K. Ludeling, M (Ed.), *Corpus Linguistics: An International Handbook* (Vol. 1, pp. 501 - 526): Walter de Gruyter.

Bar, K., Diab, M., & Hawwari, A. (2014). Arabic Multiword Expressions *Language, Culture, Computation. Computational Linguistics and Linguistics* (pp. 64-81): Springer.

Bounhas, I., & Slimani, Y. (2009). *A hybrid approach for Arabic multi-word term extraction*. Paper presented at the Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on.

Butt, M. (1999). *A grammar writer's cookbook*: CSLI.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). *Towards Best Practice for Multiword Expressions in Computational Lexicons*. Paper presented at the LREC.

Cardey, S., Chan, R., & Greenfield, P. (2006). *The development of a multilingual collocation dictionary*. Paper presented at the Proceedings of the Workshop on Multilingual Language Resources and Interoperability.

Dipper, S. (2003). *Implementing and Documenting Large-scale Grammars-German LFG*. Inst. für Maschinelle Sprachverarbeitung, Univ.

Francopoulo, G. (2013). *LMF lexical markup framework*. Hoboken, NJ; London: ISTE Ltd.

Francopoulo, G., & George, M. (2008). Language resource management-Lexical markup framework (LMF). *ISO/TC, 37*.

Francopoulo, G., & Huang, C.-R. (2014). Lexical markup framework: an ISO standard for electronic lexicons and its implications for Asian languages. *Lexicography, 1*(1), 37-51. doi:10.1007/s40607-014-0006-z

Hawwari, A., Attia, M., & Diab, M. (2014). A Framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic. *ANLP 2014*, 48.

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

Losnegaard, G. S., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S., & Monti, J. (2016). *PARSEME Survey on MWE Resources*. Paper presented at the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016).

Odijk, J. (2013). Identification and lexical representation of multiword expressions *Essential Speech and Language Technology for Dutch* (pp. 201-217): Springer.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71-106.

Rosén, V., De Smedt, K., Losnegaard, G. S., Bejček, E., Savary, A., & Osenova, P. (2016). *MWEs in Treebanks: From Survey to Guidelines*. Paper presented at the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP *Computational Linguistics and Intelligent Text Processing* (pp. 1-15): Springer.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., . . . Losnegaard, G. S. (2015). *PARSEME–PARSing and Multiword Expressions within a European* *multilingual network*. Paper presented at the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015).

Steinberger, R., Pouliquen, B., Kabadjov, M., & Van der Goot, E. (2013). JRC-Names: A freely available, highly multilingual named entity resource. *arXiv preprint arXiv:1309.6162*.

Tanabe, T., Takahashi, M., & Shudo, K. (2014). A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. *Computer Speech and Language, 28*(6), 1317-1339. doi:10.1016/j.csl.2013.09.001

Wanner, L. (1996). *Lexical functions in lexicography and natural language processing* (Vol. 31): John Benjamins Publishing.

Appendix 1. XML fragment for the MWE, *fī ʾams alḥāja*, 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 ʾams alḥāja "/>
                <feat att="Normalised Form" val="في امس الحاجة"/>
                <feat att="Different Spelling Form" val=" "/>
            </OrthographicFeatures>
```