Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure

Najet Hadj Mohamed1,2, Cherifa Ben Khelil1, Agata Savary3, Iskandar Keskes2, Jean-Yves Antoine1, Lamia Belguith Hadrich2
1 University of Tours, LIFAT, ICVL; 2 University of Sfax, MIRACL; 3 University of Paris-Saclay, LISN

Abstract
This paper describes our efforts to extend the PARSEME framework to Modern Standard Arabic. The applicability of the PARSEME guidelines was tested by measuring the inter-annotator agreement in the early annotation stage. A subset of 1,062 sentences from the Prague Arabic Dependency Treebank PADT was selected and annotated by two Arabic native speakers independently. Following their annotations, a new Arabic corpus with over 1,250 annotated VMWEs has been built. This corpus already exceeds the smallest corpora of the PARSEME suite, and enables first observations. We discuss our annotation guideline schema that shows full MWE annotation is realizable in Arabic where we get good inter-annotator agreement.

Keywords: Arabic multiword expressions, PARSEME, annotation guidelines

1. Introduction
The importance of multi word expressions (MWEs), such as by and large ‘generally’, a cheap shot ‘a cruel verbal attack’, or to eat dirt ‘to have to accept bad treatment’, has long been recognized in Natural Language Processing (NLP) (Baldwin and Kim, 2010; Constant et al., 2017). Their idiosyncratic (i.e. special and peculiar) behavior calls for language resources in which they are explicitly identified and described. Moreover, in order to enable cross-language studies of idiosyncrasy, MWE modelling should ideally be performed in a unified framework. The PARSEME community has undertaken such an effort of setting up unified annotation guidelines for verbal MWEs (VMWEs) in many languages (Savary et al., 2018; Ramisch et al., 2018; Ramisch et al., 2020). The principle of this framework is to represent in a unified way only those phenomena which are truly similar, thus emphasizing those which are specific to particular languages. So far, twenty-five national teams have prepared corpora in their languages, annotated manually for VMWEs according to the unified guidelines, and released under open licenses. This boosted the development of MWE-aware NLP tools, most prominently VMWE identifiers. Each time a new language joins PARSEME, the guidelines are tested for their applicability to this language, and modified or extended if needed. This paper describes such an ongoing effort of extending PARSEME framework to Modern Standard Arabic (MSA), henceforth called Arabic for short.

The paper is organized as follows: Section 2 describes previous works dedicated to Arabic MWEs, Section 3 is an introduction to the Arabic language from the NLP perspective, Section 4 describes linguistic properties of Arabic MWEs, Section 5 explains the construction of the Arabic VMWE-annotated corpus, including the validation and adaptation of the annotation guidelines, as well as some specific features of Arabic relevant to the annotation process. In Section 6 we report on the inter-annotator agreement at the early annotation stage. In Section 7 we present the first quantitative results of the Arabic VMWEs annotated so far and compare them to those in other languages of the PARSEME suite. Finally, in Section 8 we conclude and evoke perspectives for future work.

2. Related work
Several studies and research have been carried out on Arabic MWEs (AMWEs). Attia (2006) handles MWEs in Arabic via a finite-state machinery and the Lexical Functional Grammar (LFG). Fixed (in a nutshell) and adjacent semi-fixed (traffic light) MWEs are first processed by a composition of finite-state lexical transducers which simultaneously divides one-word phrases into components (e.g. وزیر الخارجية → and@to@minister) and joins MWEs into words with spaces (e.g. وزیر الخارجية → minister foreign ‘foreign minister’). The latter are then handled at the syntactic parsing stage as single tokens. Syntactically flexible MWEs are handled by grammar rules as syntactically compositional but as semantically non-compositional, due to the lexical selection rules. Lexical selection rules also cover phrasal verbs, e.g. a rule states that the object of rely has to be preceded by the preposition on. This shows strong links between LFG lexical rules and valence dictionaries. Attia et al. (2010) present a semi-automatic linguistic method based on regular expressions for extracting MWEs in Arabic texts. They proposed 3
approaches that focus on nominal AMWEs. The first approach relies on the correspondence asymmetries between page titles in Arabic Wikipedia and Wikipedia page titles in 21 different languages. The second approach collects English MWEs from Princeton WordNet 3.0, translates this collection into Arabic using Google Translate, and applies different search engines to validate the output. The last approach uses lexical association measures to extract MWEs from a large unannotated corpus.

Hawwari et al. (2012) created an Arabic MWE list based on a collection of 5,000 expressions manually extracted from Arabic dictionaries and grouped based on their syntactic type. Abdou (2012) explored an 83-million-word Arabic corpus in order to examine AMWEs, mainly MSA idioms, with regard to their semantics, discursive, lexical and grammatical properties. He established the main patterns of the linguistic behavior of AMWEs and developed an empirical taxonomy of six AMWE types: verb-subject, verbal, nominal, prepositional, adjectival and adverbial idioms (i.e. expressions syntactically headed by verbs, nouns, prepositions, adjectives, and adverbs, respectively). Let us take a closer look at the first two classes. A verbal idiom consists of a verb and its direct object (mostly noun) that is, at least semantically, idiomatic, e.g. 

ماذب کیم | lit. ‘he raced the wind’ ‘he ran very fast’. Verb-subject idioms are composed of one verb at least and its subject with or without any other constituents, e.g. 

عفینا نجعا | lit. ‘the glory or fame of somebody/something ended’. However, the author clarified that the verb-subject idiom term should be taken as a label only as a detailed description by itself. What is certain is that idiosyncratic combinations of verb and noun constitute a notable subclass of MWE due to their pervasiveness and their great lexical and semantic variability.

Ghoneim and Diab (2013) used LDC GALE newswire parallel Arabic-English corpus to represent MWEs in a Statistical Machine Translation (SMT) pipeline. Various types of MWEs were considered for the proposed approach: Verb-based MWEs (verb-noun constructions, verb-particle constructions, light verb constructions), Noun-based MWEs (noun-noun constructions, named entity constructions), Adjective- (AJ) and Adverb- (AV) based MWEs. A list of MWEs extracted from English WordNet database 3.0 is also used and named entities are also considered as a type of MWEs.

Al-Badrashiny et al. (2016) used a paradigm detector on the Arabic Treebank (ATB) and Arabic Gigawords corpus to build a AMWE resource. They managed to extract automatically 1,884 AMWEs. Each type of these 1,884 MWEs has 20 different forms on average due to the morphological or inflectional changes of the MWE components.

This previous work on AMWEs mainly concerned the construction of lexical and grammatical resources, as well as selected MWE-aware applications. We, conversely, focus on the construction of a MWE-annotated Arabic corpus. We chose to model AMWEs within the unified multilingual PARSEME framework (cf. Section 3). Thus, we focus not only on idioms, but also other types of VMWEs, and we test the appropriateness of the PARSEME VMWE typology for Arabic. In PARSEME, the case of Arabic is special, since efforts have already been taken towards creating an Arabic PARSEME corpus (Ramisch et al., 2018). These efforts, however, did not fully follow the PARSEME methodology, the corpus has not been openly released and is no longer available. Due to these corpus availability constraints, Arabic has never been covered by the systems developed within the PARSEME shared tasks on automatic identification of VMWEs. In order to fill this gap, we undertook the construction of a PARSEME Arabic corpus from scratch. This paper describes the first steps taken towards this aim.

3. Arabic language specificities

Modern Standard Arabic (MSA) is the universal language of the Arab world. It is a modernized and standardized version of Classical Arabic used in writing and more formal settings, such as education and media. MSA has a complex linguistic structure with a rich morphology and complex syntax (Azmi and Almajed, 2013). In this section, we give an overview of the Arabic specificities that have an impact on AMWE modelling and processing.

The Arabic language has a right-to-left writing and ambiguous shapes. It is mainly characterized by: the lack of diacritics (dedicated letters to represent short vowels), complex agglutination, pro-drop structure, and free word order structure. These characteristics make Arabic processing particularly challenging. For instance, Farghaly and Senellart (2003) estimated that the average number of ambiguities for a token in MSA can reach 19.2 (compared to 2.3 in most other languages). MSA can be fully diacritized, partially diacritized, or non diacritized. Short vowels are not often explicitly marked in writing. They are neither written in the Arabic handwriting of everyday use, nor in general publications. A non diacritized word could have different morphological features and, in some cases, different POS, especially when it is taken out of its context. In addition, even if the context is considered, the POS and the morphological features could remain ambiguous.
In addition to a concatenative morphology, where words are formed via a sequential concatenation process, the Arabic language is characterized by the presence of a templatic morphology where a morpheme is built from a root (a sequence of most often three, less so four, or very rarely five consonants), vocalisms (a collection of short vowels) and a pattern (an abstract template in which roots and vocalisms are inserted). For example, the word ‘take’ is constructed from the root أَخَذ أَخَذ أَخَذ ‘to take’, the pattern 1V2V3 and the vocalism aa (Habash, 2010). Concatenative morphemes can be stems, affixes or clitics. A clitic depends phonologically on another word or phrase and has the syntactic characteristics of a word. Clitics include prepositions, conjunctions, and pronouns. For instance, prepositions (like ل ‘for’), conjunctions (like و ‘and’) and pronouns (like ِهِ ‘he’ and مَنِّ ‘who’) can be affixed to nouns, adjectives, particles and verbs which may cause several lexical ambiguities. Indeed, the word ‘فَهم’ serves its meaning (e.g.القلوبالتيحطمها | hearts which are broken) therefore the latter is a canonical form of the former).\footnote{Pairs of words coordinated with و and which are nearly or completely synonymous}

AMWEs can be divided into nominal, verbal, adjectival, adverbial, prepositional, etc.

We have decided to focus on verbal MWEs (VMWEs) as the first step in our research and we test how well Arabic VMWEs (AVMWEs) are captured by the VMWE categories defined in PARSEME (Savary et al., 2018). This allows us to integrate Arabic into a multilingual framework, which facilitates cross-lingual linguistic studies and the development of MWE-aware tools according to common standards. For PARSEME, our work also offers the validation and extension of the unified framework to a new language.

To this aim, we firstly adopted the basic PARSEME terms, including: (i) the lexicalized components, i.e. the required and non-substitutable constituents of a MWE, (ii) the canonical form, i.e. the least syntactically marked syntactic structure of an expression which preserves its meaning (e.g.القلوبالتيحطمها | hearts which are broken) contains a plural inflection of the noun and an extraction and is therefore more syntactically marked than حَوْلَ قَلَبَاهَا | break heart her | lit. ‘he broke her heart’) then (is a canonical form of the former).\footnote{The transformation of a candidate sequence to its canonical form is necessary because the linguistic tests used in the PARSEME guidelines are syntax-driven. For instance, the headword of the canonical form of the candidate sequence is required to be a verb, which is indeed the case in broke her heart, while a non-canonical variant such as hearts which are broken is headed by a noun.} PARSEME puts forward a classification of VMWEs together with annotation guidelines for their identification and categorization. We performed pilot annotation following these guidelines, and identified the following 4 categories relevant to Arabic.

A Light Verb Construction (LVC) is formed by a (light) verb and a (predicative) noun (with an optional adposition). Its particularity lies in the fact that it is not the verb that fulfills the function of the predicate of the sentence, but rather the predicative noun. In other words, in such structures the noun expresses the action or state, while the verb carries little semantic content. The common verbs that can function as light verbs\footnote{Light verbs are called ركيزة فعلية /rakizih feeliya in Arabic (Ibrahim, 2002).} include ‘take’, قام ‘gave’, أعطى | lit. ‘he handed an advice’ ‘gave an advice’, etc. Let us consider the 2 following sentences:

(A’ata nasiha | lit. ‘he handed an advice’) ‘he gave an advice’

(A’ata kitab | lit. ‘he handed a book’) ‘he gave a book’

Both sentences are grammatically correct and they have the same structure: The verb أعطى ‘gave’ occurs
curs in past tense and the two nouns are comple-
ments of the verb. However, in the first case the
action is expressed by the verb *أعطَى* (gave), while
in the second it is expressed by the noun *نصيحة*
(advice).

PARSEME defined two subcategories for LVCs:
LVC.full (the syntactic subject of the verb is the
semantic argument of the noun) and LVC.cause (the
subject of the verb is the cause or source of the
event or state expressed by the noun). An LVC-
specific decision diagram is also provided to decide
whether a VMWE candidate should be annotated
as an LVC.full, LVC.cause or neither of the two.
Examples of these subcategories are shown in Section 4.

**Verbal Idioms (VIDs),** called *ال 활용ة الفظية* ‘م-
talazimat laafd’ya’ in Arabic, have various syntac-
tic structures and are often used in particular con-
texts and meanings. Any idiomatic expression that
has at least two lexicalised components, including
a head verb and at least one of its dependents, can
be considered a VID, as long as it does not pass the
linguistic tests for the remaining categories. Al-
though VIDs are relatively infrequent, it is noto-
riously difficult to automatically distinguish them
from literal uses of the same word combinations.
For instance *(adār lh ẓaẖ rh | lit. ‘turn his
back’)* may, depending on the context, imply that
we do not want to deal with someone or that we
are heading in the opposite direction. In fact, the
first apparent meaning of this expression is ‘to let
him behind’ but it can be also used as an idiomatic
expression meaning ‘to make him run away’ or ‘to
turn away from him’. In order to differentiate a lit-
eral from an idiomatic reading, we sometimes need
to know the morpho-syntactic variations allowed or
prohibited by a VID. For instance, in *(aẖḏ bīdh | took with+hand+his | lit. ‘took his
hand’)* ‘gave a helping hand’ the noun is agglutinated with
a possessive and the preposition *ب* (together with) ‘be patient and you’ll get what you want’.
The last step in adapting the PARSEME guidelines
to Arabic was the insertion of (dozens of) Arabic
examples, mostly stemming from the pilot annota-
tion, as illustrations of the VMWE categories and
linguistic tests. Figure 2 shows a sample test with
elements displayed for Arabic and English.

5. Corpus construction

Once the annotation guidelines were tested against
sample texts in pilot annotation, we proceeded to
systematic annotation of a pre-existing syntacti-

5.1. Source Corpus

The PARSEME format builds upon morphosyn-
tactic annotation in the CoNLL-U format 6, which
is a de facto standard for dependency annotation
defined by Universal Dependencies (UD). Since,
additionally, our Arabic corpus is to be released
openly, we chose the only Arabic UD corpus whose
data are fully openly available, i.e. the Prague
Arabic Dependency Treebank (PADT) [Hajic et al.
2009]. It was initially created as a multi-layer de-
pendency treebank, with a morphological, an an-
alytical (surface-syntax) and a tectogrammatical
(deep-syntax) layer, and further converted to the
CoNLL-U format. It currently has 7,664 annotated
sentences (282,384 tokens) from newswire sources.

---

6This category is easily mistaken for a Verb Particle
Construction, also defined in the PARSEME typology.
In Arabic, particle added to a verb can be either a pro-
clitic or a preposition, thus, the term particle is often
used to denote both. The appropriate term here is the
preposition since a verb with its preposition will always
be followed by a noun phrase otherwise the sentence
will be incomplete.

6https://universaldependencies.org/format.html
5.2. VMWE Annotations

The PARSEME annotation guidelines are conceived so as to ensure reproducibility of the VMWE identification and categorization. Namely, they are organized as decision diagrams based on linguistic tests. As long as the answers to the atomic tests are the same, the final annotation decision remains stable.

Based on these guidelines, we manually annotated VMWE occurrences in PADT. The guidelines make us proceed as follows. Firstly, we identify a candidate, that is, a combination of a verb with at least one other word which could form a VMWE. The candidate is then transformed to its canonical form (cf. Section 4), and the subsequent tests are applied to this form. Secondly, we determine the lexicalized components (cf. Section 4). Thirdly, we apply tests according to the decision trees of the candidate’s possible category (e.g. Figure 1 represents the LVC-specific decision tree). After these tests, we are able to decide whether the candidate is indeed a VMWE, and, if so, what is its category. Finally we apply the (optional) tests for IAVs.

To illustrate these steps, let us consider the following examples:

(a) أسدى النصيحة (asda al-naṣīḥẗ | lit. ‘he weaves an advice’) ‘he gives an advice’
(b) يعرض للخطر (īuʿariḍu lilẖaṭar | lit. ‘offers to the danger’) ‘exposes someone or something to danger’
(c) وضعه على الرف (waḍaʿahu ʿalai al-rwaf | lit. ‘put it on the shelf’) ‘put it aside or ignore it’
(d) شهيد إعصار (šahīda iʿṣār | lit. ‘witness hurricane’) ‘experience a hurricane’

Figure 1: LVC-specific decision tree.

Each of these examples contains a head verb and a complement, possibly introduced by an adposition. In such a case, the global decision diagram (not shown here), first redirects to the LVC-specific decision tree (in Fig. 1). Example (a) passes the first 4 tests: The noun النصيحة ‘advice’ is abstract (LVC.0) and predicative (LVC.1); the subject of the verb is a semantic argument of the noun (LVC.2); the verb أسدى ‘gives’ adds no meaning to the noun (it only expresses performing the activity of the noun, LVC.3) and lastly the verb reduction yields an NP referring to same event as the one expressed by the noun (LVC.4).

As for (b), the noun خطر ‘danger’ is abstract (LVC.0) and predicative (LVC.1), but it does not pass test LVC.2. Figure 2 contains an excerpt with some examples to illustrate how to use test LVC.2. In (b), the subject of the verb (he) is not the agent of the noun خطر ‘danger’ but rather the cause of the predicate expressed by the noun. In other words, the subject of the verb represents the source of the state ‘to be exposed to danger’ referred to by the noun. Thus, (b) passes test LVC.5 and is annotated as an LVC.cause.

Example (c) fails test LVC.0 since the noun الرف ‘the shelf’ is not abstract, therefore it is not an LVC, so we go on with the VID decision tree. There, (unlike in the LVC tree) the candidate needs to pass at least one inflexibility test. Example (c) exhibits lexical inflexibility: If we replace one of its components, in this case الرف ‘the shelf’, by a semantically related word (taken from a relatively large semantic class), such as الطاولة ‘the table’, we loose the idiomatic sense of the expression.

The last example (d) does not pass test LVC.2 since the noun إعصار ‘hurricane’ has no semantic argument. It also fails all the VID tests. Therefore, it is neither LVC nor VID and cannot be considered a VMWE.

Figure 2: Excerpt from PARSEME guidelines: Test LVC.2 - Verb’s subject is noun’s semantic argument?

5.3. Challenging phenomena

The manual annotation task brought us its own set of challenges. We encountered various issues,
have several possible segmentations, we had to choose the right lexicalized components of the expression carefully. Consider the AVMVE

وضع به (۶عُدَّ) ۶لِي ۶الرف [lit. ‘put him on the shelf’] ‘ignored him’. Here, the agglutination of the verb (۶عُدَّ) ‘to put’ to the enclitic (۶ة) ‘him’ is required. Furthermore, if we do not count the enclitic as a lexicalized component, we lose the idiomatic sense of the expression and it becomes a literal expression meaning ‘to put something on the shelf’.

Consider another example:

الضروري إجراء وнестиح (۶هد ۶الإجرا ۶ودرِى) ۶will and take the action necessary ‘he will take the necessary action’. Here, the verb (۶هد) ‘to take’ is agglutinated to 2 proclitics: ۶ة (‘and’) indicates coordination, while ۶س ‘will’ indicates an action in the future. Also, the noun (۶يجرا) ‘procedure/action’ is agglutinated to the enclitic (۶ة) ‘the’ which has the role of a determiner. In this case the agglutinated morphemes are not required for the idiomatic meaning to occur, so we select only the lexicalized components without agglutination (۶هد إجراء ۶ية إجراء | lit. ‘take an action’).

Masdar is a specific Arabic part of speech defined as a noun of the verb, which expresses the same event as the corresponding verb stem, but without any reference of time. According to the PARSEME guidelines, meaning-preserving morphosyntactic variants of a VMWEs should be annotated. Therefore, if a verb in a VMWE is replaced by its masdar and the idiomatic meaning is kept, we consider that this is a valid occurrence of a VMWE like in the following example:

على المجتمع تحقيق نهضته (۶الإمساك ۶تحقيق ۶نهضه) ‘the society must achieve its renaissance’ ‘the society must achieve its renaissance’.

The meaning is carried by the noun (۶نهضه) ‘renaissance’, while the masdar (۶تحقيق | lit. ‘achievement’) ‘realization’ derived from the verb (۶حققه) ‘realize’ behaves as a light verb. In this case, the candidate VMWE occurrence is (۶تحقيق نهضه) ‘achievement’ (lit. ‘achievement renaissance’) ‘achievement of renaissance’, and the canonical form to which the linguistic tests are applied is (۶حققه نهضه) ‘make a renaissance’ which passes the LVC.full tests.

An interesting specificity of Arabic is the largely productive syntactic pattern (light verb + Masdar) where the verb and the Masdar are derived from the same verbal root, which leads to a semantic duality, like in the following example:

خراج هرجا (۶رفع هرجا | lit. ‘he exited exit’) ‘he went out’. Such verb/masdar combinations pass the LVC.full tests and are annotated as such.

MWE categorisation challenge A major disagreement among annotators concerned the distinction between LVC and VID for candidates con-
sisting of a verb and a noun, where the noun is the
direct object of the verb.
A VID candidate, can be read both literally and
idiomatically in two different contexts. For example,
(qāṭ al-ṭrīq mosrʿ | lit. ‘cut road
rushing’) ‘rush across the street’ is a literal ex-
pression referring to the action of crossing the street.
Conversely, (qāṭ al-ṭrīq `lh | lit. ‘cut
the+road on+him’) ‘cut off his road’ is idiomatic,
meaning ‘to prevent someone from doing what he
wants to do’. In such cases, the annotator should
strive to fully understand the context of the ex-
pression and decide if it is indeed a VID.
The second disagreement requires a double effort
from the annotators while putting the expression
in several contexts so that he can judge the type.
For example, (škl ġzʿ | lit. ‘shape part’)
‘be part of’ is a VID and not an L VC because it
does not allow the verb to be omitted, although
the noun is predicative and keeps its usual sense.
Conversely, in the LVC.full (أسدى النصيحة
(āṣda al-anṣāḥ | lit. ‘he weaves an advice’) ‘he gives an
advice’, the verb أسدى ‘to weave’ is semantically
rich in other contexts but in this expression it acts
as the light verb equivalent to أصلح (a’ta) ‘to give’,
and can be reduced as required by test LVC.4 (cf.
Fig. 2).

6. Inter-Annotator Agreement
The previous sections described the process of
adapting the pre-exiting PARSEME VMWE anno-
tation methodology and guidelines to Arabic. To
qualitatively assess the reliability of this adapta-
tion, we measured the inter-annotator agreement
in the early annotation stage. A subset of 1,062
sentences from the PADT corpus was selected and
notated by two Arabic native speakers independ-
ently.

|      | A₁ | A₂ | F_span | κ_span | κ_cat |
|------|----|----|--------|--------|-------|
|      | 763| 704| 0.699  | 0.626  | 0.864 |

Table 1: Inter-annotator agreement on a sample of 1062 sentences, with A₁ and A₂ VMWEs anno-
tated by each annotator. F_span is the F-measure
between annotators, κ_span is the agreement on the
annotation span and κ_cat is the agreement on the
VMWE category.

Table 2 shows the inter-annotator agreement (IAA) calculated with the PARSEME tool²⁴. An-
otators A₁ and A₂ annotated 763 and 704 occur-
rences of VMWEs, respectively. Their agreement is
measured separately for unifying (i.e. identifying
the appropriate text spans) and for categoriza-
As discussed by (Ramisch et al., 2018), F_span
is the MWE-based F-measure of A₁’s annotations
with respect to A₂, and vice versa. With this mea-
sure, an annotation is considered correct if both
annotators identified precisely the same tokens as
belonging to one VMWE (i.e. partial overlaps are
considered fully erroneous). Then, κ_span estimates
to what extent the observed agreement P₀ exceeds
the expected agreement Pₑ, that is, κ = P₀−Pₑ
. The expected agreement Pₑ is approximated by
the number of verbs in the text (a VMWE usually
contains a verbal head). Finally, κ_cat is calculated
on those VMWEs for which both annotators agree
on the precise spans.

We compared these initial IAA scores for Arabic
to those of the PARSEME suite (editions 1.1 and
1.2)²⁵ Among the 26 IAA estimations²⁶ Arabic
now has:

- the second highest (after Chinese) number of
VMWE annotations used to estimate the IAA,
- the 12th, 14th and 12th best F_span, κ_span, and
κ_cat, respectively.

Note that, for the other languages, the IAA corpus
was often double-annotated at the final stage of the
annotation campaign, when annotators’ expertise
has reached its optimum. The IAA for Arabic,
conversely, was estimated at the early annotation
stage, so as to control the sufficient preparedness of
the annotators and the soundness of methodology
at its start. Thus, not only does Arabic already
have state-of-the-art IAA scores, but its consis-
tency is expected to grow, notably when the final
PARSEME consistency check procedures (Savary
et al., 2018) have been applied.

7. Corpus Analysis
Table 3 gives that statistics of the Arabic corpus in
its current state (considering the annotator who
identified the highest number of VMWEs). Its
size, with over 1,250 annotated VMWEs, already
exceeds the smallest corpora of the PARSEME
suite, and enables first observations. The density
of VMWEs is of about 0.68 VMWEs per sentence.
The universality (i.e. presence in all the languages
under study) of the VID and LVC categories is con-
firmed, with LVC.full being almost twice as fre-
quent than VID, and LVC.cause being sporadic.

²⁴Like for other languages, this approximation is
slightly biased by the syntactic variants in which nom-
inalizations or participles derived from verbs might be
notated as nouns or adjectives. Here, notably mas-
dar variants are concerned (cf. section 5.3).
²⁵Edition 1.0 is not considered here since it was based
on a different version of the annotation guidelines.
²⁶For 3 languages in the PARSEME suite the IAA
was estimated twice: once in edition 1.1 and once in
1.2. We neglect the previous publicly unavailable Ara-
bic PARSEME corpus.
The quasi-universal IRV and VPC categories (frequent in Romance and Slavic languages, as well as in German, but absent or rare in other languages) are not found in Arabic.

Following, Savary et al., 2018, we also analysed the corpus in terms of lengths (i.e. numbers of tokens) of the annotated VMWEs and of their discontinuities (i.e. numbers of external tokens occurring between the first and the last token of a VMWE). Discontinuities are a major challenge for the MWE identification task (Constant et al., 2017), therefore their distribution is an important feature of the language and of the corpus under study. Table 2 shows the results of this analysis. In particular, over 73% of all VMWEs contain 2 tokens (column 3), above 42% are continuous (column 7) but more than 17% (last column) have more than 3 gaps.

We compare these results to the 18 languages from the PARSEME suite in edition 1.0. The average length of Arabic VMWEs (2.26 in column 1 of Table 2) is not an outlier, since in 17 (out of the 18) languages this factor is between 2.02 and 2.71. The non-existence or rareness of single-token VMWEs (0 in column 2 of Table 2) is also a feature of 14 languages (Hungarian, German and Portuguese being outliers in this category). In terms of discontinuities, Arabic is the second most outstanding language (after German). It has 1.97 gaps on average (German has 2.96, Slovene 1.47, Czech 1.35, Hungarian 1.01, and all other languages have less than 1). It also has the 2nd lowest percentage of continuous VWEs (42.17%) and the 2nd highest percentage of VMWEs with gaps longer than 3 (17.33%), after German (35.7% and 30.5%, respectively).

The corpus in its current state is already available in the PARSEME repository under the CC-BY v4 license, including the double-annotated IAA sample. Thus, the results presented here are fully reproducible, using the PARSEME utilities.

8. Conclusion and Future Work

The main contribution of this work is to create an openly accessible Arabic corpus enriched by VMWE annotations. For this, we manually annotated the PADT corpus using the PARSEME guidelines, which have shown their effectiveness on MSA. The VMWE types occurring in Arabic are verbal idioms, light verb constructions, multi-verb constructions inherently adpositional verbs. These phenomena were annotated on a sample of 1,062 sentences from the PADT corpus by 2 annotators and we get reasonable inter-annotator agreement. We have annotated 1,252 VMWE occurrences with a high rate of discontinuous expressions (58%).

We explained challenging phenomena, stemming from the rich and complex morphology, as well as to non-vocalized spelling, notably high level of ambiguity, agglutination, discontinuities and morphosyntactic variation.

Since PARSEME guidelines proved perfectly adaptable to MSA, we consider the initial initiative of the annotation task as validated. However, it is possible that we missed some types of variation not represented in our corpus. Ongoing work consists in annotating texts from new sources and genres. This might trigger Arabic-specific additions to the guidelines. We are also in the process of training MWE identifiers on the corpus, these results should be published soon.

References

Abdou, A. (2012). Arabic Idioms: A Corpus Based Study. Routledge.

Al-Badrashiny, M., Hawwari, A., Ghoneim, M., and Diab, M. (2016). SAMER: a semi-automatically created lexical resource for Arabic verbal multiword expressions tokens paradigm.

Table 2: Statistics of the Arabic VMWE corpus in its current state, in terms of the number of sentences and tokens, as well as the number of annotated VMWEs per category and in total.

| # Sent. | # Tok. | VID | IRV | LVC.full | LVC.cause | VPC.full | VPC.semi | IAV | MVC | All |
|--------|--------|-----|-----|----------|-----------|----------|----------|-----|-----|-----|
| 1,847  | 70,498 | 365 | 0   | 673      | 68        | 0        | 0        | 185 | 1   | 1,252 |

12 This is in contrast with the statistics of the previous PARSEME Arabic corpus, which we could not recalculate due to the unavailability of this corpus. There, 4,219 VMWEs were reported in 3,137 sentences (with the density of 1.35) split into: 1,769 LVCs.full, 1,320 VID.s, 1,080 VPCs, 17 IRVs, 33 MVC and 0 LVC.cause. Note in particular the absence of VPCs in our statistics. We claim that particles, as defined in PARSEME, are non-existant or very rare in Arabic. The VPCs from the Hawwari corpus might likely be IAVs.

13 The statistics from the following editions have not been published.

14 Note a token can contain several agglutinated words, so it can, indeed, be a MWEs.
and their morphosyntactic features. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12), pages 113–122.

Attia, M., Toral, A., Tounsi, L., Pecina, P., and Van Genabith, J. (2010). Automatic extraction of Arabic multiword expressions. In Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, pages 19–27.

Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In International Conference on Natural Language Processing (in Finland), pages 87–98. Springer.

Azmi, A. M. and Almajed, R. S. (2015). A survey of automatic arabic diacritization techniques. Natural Language Engineering, 21(3):477–495.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Nitin Indurkhya et al., editors, Handbook of Natural Language Processing, Second Edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Survey: Multiword expression processing: A Survey. Computational Linguistics, 43(4):837–892, December.

Farhaly, A. and Senellart, J. (2003). Inductive coding of the Arabic lexicon. In Workshop on Machine Translation for Semitic languages: issues and approaches, New Orleans, USA, September 23-27.

Ghoneim, M. and Diab, M. (2013). Multiword expressions in the context of statistical machine translation. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 1181–1187, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Habash. (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1):1–187.

K. (2009). Prague arabic dependency treebank 1.0.

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, pages 24–29.

Ibrahim, A. H. (2002). Les verbes supports en arabe. Bulletin de la Société de Linguistique de Paris, 97(1):315–352.

Koch, F. and Wieser, W. (1983). Partitioning of Energy in Fish: can Reduction of Swimming Activity Compensate for the Cost of Production? Journal of Experimental Biology, 107(1):141–146, 11.

Maamouri and Bies. (2010). Arabic Treebank: Part 3 v 3.2 LDC2010T08. Web Download. Philadelphia: Linguistic Data Consortium.

Othman, J., Bennett, J., and Blamey, R. (2004). Environmental values and resource management options: a choice modelling experience in malaysia. Environment and Development Economics, 9(6):803–824.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Inurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartin, C., Qasemizadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 222–240, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118. Association for Computational Linguistics, December.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaite, J., Krek, S., Liebeschkind, C., Monti, J., Escartin, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, et al., editors, Multitword expressions at length and in depth: Extended papers from the MWE 2017 workshop, pages 87–147. Language Science Press, Berlin.