PARSEME-AR: Arabic reference corpus for multiword expressions using PARSEME annotation guidelines

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskander Keskes, Jean Yves Antoine, Lamia Belguith Hadrich

PARSEME-AR: Arabic reference corpus for Multiword Expressions using PARSEME annotation guidelines

Najet Hadj Mohamed[1,2*], Cherifa Ben Khelil[4,2], Agata Savary[3], Iskander Keskes[1], Jean Yves Antoine[2] and Lamia Belguith Hadrich[1]

[1]MIRACL, University of Sfax, Sfax, 3029, Tunisia.
[2]LIFAT, ICVL, University of Tours, Tours, 37000 , France.
[3]LISN, University of Paris-Saclay, CNRS, Orsay, 91400, France.
[4]EFREI Research Lab , University of Paris Panthéon Assas, Villejuif, 94800, France.

*Corresponding author(s). E-mail(s): najat.hadjmohamed@etu.univ-tours.fr;
Contributing authors: cherifa.bk@gmail.com; agata.savary@universite-paris-saclay.fr; iskandar.keskes@fsegs.usf.tn; Jean-Yves.Antoine@univ-tours.fr; lamia.belguith@fsegs.usf.tn;

Abstract

In this paper we present PARSEME-AR, the first openly available Arabic corpus manually annotated for Verbal Multiword Expressions (VMWEs). The annotation process is carried out based on guidelines put forward by PARSEME, a multilingual project for more than 26 languages. The corpus contains 4,749 VMWEs in about 7,500 sentences taken from the Prague Arabic Dependency Treebank. The results notably show a high degree of discontinuity in Arabic VMWEs in comparison to other languages in the PARSEME suite. We also propose analyses of interesting and challenging phenomena encountered during the annotation process. Moreover, we offer the first benchmark for the VMWE identification task in Arabic, by training two state-of-the-art systems, on our Arabic data.

Keywords: Arabic, Multiword expressions, PARSEME, annotation guidelines

1 Introduction

Multiword Expressions (MWEs) currently represent an object of study common to many disciplines in the science of language. Along with simple words, these units are part of the lexical component of each language. Due to their non-compositional and prefabricated nature, they exhibit particular linguistic specificities which distinguish them from literal units. These specificities lie in the lexical, morphosyntactic, and semantic idiosyncrasies of MWEs, which call for diverse linguistic analyses. Due to the pervasiveness of MWEs, their detection is a critical issue. For instance 41% of WordNet are MWEs (Elkateb et al, 2006). Jackendoff (1997) estimates that each language contains as many MWEs as isolated words. Moreover, it is impossible to establish an exhaustive list of MWEs because new expressions are constantly appearing, whether they are polylexical named entities (names of people, organizations, etc.) or neologisms obtained by reusing the available lexicon (Gross, 1982). Baldwin and Kim (2010) mention that 8% of parsing errors are caused by the lack of MWE detection. Numerous studies devoted to automatic detection of MWEs are justified by their evolutionary nature.

The PARSEME (PARSing and Multi-word Expressions) network (Savary et al, 2017b) is focused on MWEs. PARSEME brings together experts (linguists, computer scientists, computer linguists, etc.) from different countries with the goal of improving the quality of natural language processing (NLP) systems by taking MWEs into account. This network focused on refining the annotation methodology for verbal MWEs (VMWEs) as well as creating annotated corpora. So far, twenty-six 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 an effort to extend the PARSEME framework to Modern Standard Arabic (MSA), henceforth called Arabic for short. We build upon our previous work described in Hadj Mohamed et al (2022). We largely extend and update those previous results by: (i) a larger corpus, (ii) a more detailed description of the corpus construction process, (iii) more in-depth corpus analyses, (iv) training state-of-the art MWE identifiers on the final corpus, (v) discussions on their results, i.e. the first results – to our knowledge – of the MWE identification task ever published for Arabic, (vi) error analysis and first steps towards future enhancements.

The paper is organized as follows: Section 2 is an introduction to the Arabic language. Sections 3 and 4 describe linguistic properties of MWEs in general, and of various categories of VMWEs in Arabic, as defined in the PARSEME guidelines. Section 5 explains the construction of the Arabic VMWE-annotated corpus and its quality estimation. In Section 6 we present the quantitative results of the Arabic VMWEs annotations. In Section 7 we offer analyses of

some outstanding linguistic phenomena encountered during the corpus annotation process. In Section 8 we present the results of a benchmark for the MWE identification task and an analysis of error committed by the two state-of-the-art systems trained on our corpus. Section 9 describes previous works dedicated to Arabic MWEs and compares our contribution to this state of art. Finally in Section 10 concludes and evokes perspectives for future work.

2 Arabic language

Today, the term "Arabic language" can refer to MSA or a number of spoken vernaculars known as Arabic dialects. Unlike dialects, that result from linguistic interference between the Arabic language and local or neighboring languages, MSA represents a modernized and standardized version of classical Arabic. It is used in writing and in more formal settings, such as literature, education, academia, and media. MSA is thus viewed as the universal language of the Arabic world. As mentioned above, our research focuses on MSA. Therefore, we dedicate this section to highlighting some important morpho-syntactic specificities of MSA.

First of all, it should be noted that the Arabic language is a Semitic language that has a right-to-left writing with 28 letters. Most of these letters are clumped together when written and they can be joined by various long and short vowels. Arabic documents and texts can be fully vocalized (e.g. Qur'anic texts, textbooks, children's stories), partially vocalized (e.g. books) or simply not vocalized (e.g. news wire). It is more common to find not vocalized texts. Therefore, when these texts are analyzed, the degree of ambiguity may be very high. Farghaly and Senellart (2003) assessed that the average of ambiguities for a non vowelised token in MSA can reach 19.2. For instance, the unvoweled word جرح (ğrḥh/'to hurt') can have the following vowelled versions: جَرَحَ (ğārāḥā/ 'he hurt'), جُرِحَ (ğuriḥa/ 'he was wounded') and جُرْحٌ (ğurohḥ/ 'injury'). This example demonstrates that a word without a vowel can have distinct POS and morphological features, especially when it is taken out of its context. Even when the context is considered, the POS and the morphological features might remain ambiguous. Let us see the unvoweled sentence (1).

(1) مجموعة من الأدوية لمعالجة ألمه و جرحه
mğmūʿT mn al-a'dwyT lmʿālğT al-mh ū ğrḣh
the-series from the-medicaments to-treat pain-his and injury-his
وصف الطبيب
ūṡf al-ṫbyb
describes.3.SG.PAST the-doctor

a. 'The doctor described a series of medications to treat his pain and hurt him'

b. 'The doctor described a series of medications to treat his pain and injury'

The leftmost word جرحه (ğrḣh) can be interpreted either as a noun جُرْح (ğuroḣ/ 'injury') with an agglutinated possessive ه (h/ 'his') or as a verb جَرَحَ (ğārāḣā/ 'he hurt') with an agglutinated anaphoric proun ه (h/ 'him'). Even if Arabic-speaking readers will tend to think that the second interpretation (b) is more likely to be the right one, only the remainder of the sentence (or paragraph) will confirm this.

Besides, as highlighted in this example, Arabic is an agglutinating language. Agglutination is the process of adjoining clitics to simple word forms to create more complex forms. These clitics include prepositions (e.g. ل 'for'), conjunctions (e.g. و 'and'), articles (e.g. اَل 'the') and pronouns (e.g. ه 'he'). They can be affixed to nouns, adjectives and verbs which may cause several lexical ambiguities. For example, the word وهم can be interpreted as the noun وَهْمٌ (wahmo/ 'illusion'), as the verb وَهَمَ (wahama/'to imagine') or also as the conjunction وَ (wa/ 'and') followed by the pronoun هُمْ (hum/ 'they'). Boudchiche and Mazroui (2015) make a statistical study on the level of ambiguity caused by the absence of diacritical marks in Arabic texts, using 80 million Arabic words. The percentage of non-diacritical words that have more than one potential root is equal to 28.47%, increasing to 75.08% when considering the lemma.

Arabic has a relatively free word order. Indeed, it is possible to change the order of the words in the sentence without altering its meaning. Only a few cases of changes require the application of new agreement rules so that the sentence remains grammatically correct and retains the same meaning. Namely, the most frequent orders in MSA for a sentence are VSO (Verb, Subject, Object) (Benmamoun, 1997) and SVO (Lyovin et al, 1997). In SVO, the verb agrees with the subject in number, gender, and person, while in VSO, the verb partially agrees with the subject (in gender and person, but not in number). OVS and VOS are two other possible orders, but one can also find in practice OV and VO sentences. MSA has a pro-drop specificity i.e. dropping

the subject pronoun. The subject is implied, this makes it an elliptical clause. In the following example, ركب السيارة 'ride the-car', the verb is conjugated in the third-person singular past tense and its subject is elliptical.
While our paper primarily focuses on the development of MWE identification tools for MSA, it is important to consider the broader implications for other Arabic dialects. MSA serves as a standardized form of Arabic used in formal and written contexts across various Arabic-speaking regions. However, the extent to which resources developed for MSA can be applied to other Arabic dialects remains a significant question. This issue raises considerations regarding the representativeness of MSA and the challenges inherent in adapting MWE identification tools to diverse dialectal variants.

3 Multiword Expressions

Following the PARSEME approach (Savary et al, 2018a), MWEs are understood in this work as combinations of two or more words which display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy, formalised by the annotation procedure. Components of a MWE which are always realized by the same lexemes are called *lexicalized components*. Henceforth, they will be called components for short. They are highlighted in bold in the examples that are provided in this paper. The following is how Savary et al (2018a) interpret this idea of component lexicalization:
lexicalisation is traditionally defined as a diachronic process by which a word or a phrase acquires the status of an autonomous lexical unit, that is, a form which it could not have if it had arisen by the application of productive rules. (...) Our notion of lexicalisation extends this standard terminology, as it applies not only to VMWEs but to their components as well.
In this way, all of the expressions in our work, that are regarded as MWEs, are lexicalized, and their fixed components are also called lexicalized. The particular aspect of Arabic is that, due to agglutination (cf. Section 2), words often do not coincide with graphical tokens. Thus, a graphical token can include several MWE components, as in example (2), where 3 graphical tokens cover 5 words. Moreover, as in example (3), one graphical token, here: **القرار** (**qrAr**-al / decision-the / lit. 'the decision') , can contain several words among which some are and some are not lexicalized components of an MWE.

(2) ركبه على ملحه
rkbt-h **‘la** **mlḥ-o**
knees-his on salt-his
lit. his salt on his knees / ‘someone who gets angry easily’

(3) القرار أخذ
qrAr-al **aẋḍ**
decision-the take.3.SG.PAST
lit. he took the decision / ‘he decided’

Many works focus on the challenging behaviour of MWEs and have been reported to improve NLP tasks, such as syntactic parsing (Nivre et al, 2004), machine translation (Deksne et al, 2008) and text mining (SanJuan and Ibekwe-SanJuan, 2006). The major characteristic of MWEs is that they exhibit idiosyncratic (unusual) behavior such as the inability to predict the properties of the expression from those of its components. This is known as the non-compositionality principle, which is demonstrated by the next paragraphs.

Morphosyntactic inflexibility: Morphosyntactic inflexibility is a key feature of some MWEs (Svensson, 2004). It refers to the morphological and syntactic phenomenon in which a potential word form of a MWE component cannot surface without the loss of the idiomatic reading. This can happen for various features like the number, gender, tense, aspect, etc. In English, the expression *to* ***turn turtle*** ‘to turn upside down’ is accepted by a native speaker, but not *to turn turtles*, etc. This is also true in Arabic. Speakers use the Arabic expression (4) but not (5). Pluralizing the word الفرس (‘the-horse’) changes the meaning of the expression and it is no longer considered as a MWE.

(4) الفرس مربط
al-frs **mrbṭ**
the-horse stall
lit. the stall of the horse / ‘moral of the story’

(5) الفروس مربط
al-frus mrbṭ
the+horses stall
lit. the stall of the horses

In some expressions, the opposite is impossible. The plural form of a word cannot be changed to the singular, as shown in the following example (6). The expression loses its idiomatic meaning if we replace الكلاب (‘the-dogs’) by الكلب (‘the-dog’).

(6) تسير القافلة و تنبح الكلاب
tsīr al-qāflT ū tnbḥ al-klāb
passes the-caravan and bark the-dogs
lit. the dogs bark and the caravan passes / ‘life goes on even if some will try to stop progress’

Syntactic blocking can be related to passivation, pronominalization, interrogation, relativation, and other syntactic transformations. Example (7) is an idiom in the active voice, but, the idiomatic meaning is lost if the expression is used in the passive voice (cf. example 8).

(7) أخيه لحم أكل
axīḥ+h **lḥm akl**
brother.GENITIVE+his flesh eat.3.SG.PAST
lit. he ate his brother’s flesh / ‘he talked badly behind someone’s back’

(8) مأكول أخيه لحم
m’akūl aḥī+h lḥm
was.eaten brother.GENITIVE+his meat
lit. the meat of his brother was eaten

Lexical inflexibility: This property means that a frozen expression would be distorted if a word were replaced by a synonym, a hyperonym, or a word from the same semantic class. In example (9), if we replace the term تفاحة (‘apple’) by عنب (‘grape’) or أدم (‘Adam’) by أحمد (‘Ahmed’), the idiomatic character of the expression would be lost.

(9) أدم تفاحة
adm tfāḥT
Adam apple
lit. Adam’s apple / ‘laryngeal mound’

Semantic non-compositionality: This term names the fact that a MWE’s meaning cannot be straightforwardly inferred from its parts. Semantic non-compositionality is a matter of scale and PARSEME claims that it is hard to test it directly. Instead, it can be approximated by morphosyntactic and lexical inflexibility. Thus, in compositional expressions, we can substitute one of the components by some of its synonyms while maintaining the idiomatic meaning, as in the example (10), where replacing أحرزت (’made.she’) by قامت (’made.she’) will not change the meaning of the expression. On the other side of the compositionality scale, completely opaque and idiomatic expressions do

not license any morphosyntactic alternations, as illustrated in the example (11). Here, replacing the verb أخذ ('took') by أمسك ('catch') will only allow for the literal meaning.

(10) أحرزت تقدم
Aḥrzt tqdm'
make.3.SG.PAST progress
lit. she made a progress 'she progressed'

(11) أخذ الثور من قرنيه
axḍ al-ūr mn qrnī+h
take.3.SG.PAST the-bull from horns-his
lit. he took the bull by its corns / 'he faced hard situation with courage'

Pragmatic Idiosyncrasy: This term refers to the situation in which a MWE can convey a particular meaning in a specific enunciation context, like in (12) which is an expression often used as a prayer for the dead.

(12) رحمه الله
rḥm-h al-lh
mercy+him God
lit. God mercy him / 'May God have mercy on him'

4 Verbal Multiword Expressions

Following PARSEME, our study primarily focused on VMWEs. A VMWE is defined as an MWE whose canonical form (the least syntactically marked variant) has a verb as its syntactic head, and whose distribution is that of a verb, a verb phrase or a sentence.

Among the VMWE categories defined by PARSEME, four are relevant to Arabic.

1. **Verbal Idioms (VIDs)** are idiomatic expressions that consist of at least two lexicalized components, including a head verb and at least one of its dependents. They have a variety of syntactic structures and are frequently used with specific meanings and contexts. Their meaning is not compositional and they often have a double literal/idiomatic reading as in (13). The mental image and the meaning of the expression are closely related when they are transparent, as in the example (13). In contrast, when they are opaque, no clear link is made between the mental image and the meaning of the VID as in the example (14).

   (13) طوى الصفحة (category: *VID*)
   :twi al-ṡfḥT
   turn.3.SG.PAST page-the
   lit. he turned the page / 'he started something new'

(14) فَأْرًا فولد الجبل تَمَخَّض (category: *VID*)
fa'arā f-ūld al-ǧbl tamaẋaḍ
mouse so-it.beget the-mountain labored
lit. the mountain labored so it beget a mouse / 'someone talks a lot but hardly acts'

2. **Light Verb Constructions (LVCs)** are verb-noun constructions (sometimes with an additional preposition) in which the noun carries the semantic meaning of the expression while the verb is semantically weak. The verb is classified as a light verb and it only conveys personal and temporal inflection information. For instance, rather than employing the LVC expression (15), we can simply say مهم أحمد دور ('Ahmed's role is important'). PARSEME subdivided the LVCs into 2 subcategories: 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). In an LVC.full, the verb must not be aspectual (begin, continue, cease, finish). Furthermore, we are not restricted to frequently meaningless verbs like أخذ 'take', أعطى 'give', etc. PARSEME considers a verb as a light verb if the noun is self-sufficient as a bearer of meaning, as in the example (16). Even though the verb أسدى 'weave' has its own semantics as a standalone verb, here it is considered as light verb because the expression's meaning is carried solely by the noun. An example of LVC.cause is (17). Here, the verb أعطى 'give' adds causative meaning to the noun.

(15) مهم دورا أحمد لعب (category: *LVC.full*)
mhm **dūrā** Ahmed **l'b**
important role Ahmed played
lit. Ahmed played an important role / 'Ahmed's role was important'

(16) نصيحة أسدت (category: *LVC.full*)
nsīḥT asda-t
advice weave.3.SG.PAST
lit. she wove advice / 'she gave advice'

(17) اعطت حق (category: *LVC.cause*)
a'a-t **hq**
give.3.SG.PAST right
lit. she gave a right / she granted the right'

3. **Multi-verb constructions (MVCs)** are expressions consisting of two adjacent verbs, which display lexical inflexibility, such as in (18)

(18) تستطيع حاول (category: *MVC*)
tstṫī' **hāūl**
can.2.SG.PRESENT try.2.SG.IMPERATIVE
lit. you try you can, / 'if you try you can reach what you want'

4. **Inherently Adpositional Verbs (IAV)** consist of a verb and an idiomatically chosen adposition, such as a preposition or postposition. The preposition is either always required or, if absent, changes the meaning of the verb of VMWE significantly. For example, the verb اشاد (ashād) can be translated as ('to raise'), as in e.g. أشاد المبنى (ashād al-bnā) 'he raises the construction'. However, when associated with the preposition ب (bi) 'for/with', it indicates praising someone or something as in the example (19).

(19) اشاد ب العمل (category: *IAV*)
ašād **bi** al-aml
3.SG.PAST with raise. work
lit. he raised with the work / 'she paid tribute to the work'

The linguistic tests for IAVs proved partly satisfactory only, therefore IAV is considered an experimental category and is annotated optionally in PARSEME. We do annotate IAVs in Arabic (also experimentally), notably so as to discuss terminological issues on the border with verb-particle constructions, as already mentioned in related work on Arabic (cf. Section 9).

5 Corpus Construction

This section discusses the various steps we took to create our VMWE-annotated corpus, as well as the data and tools we used.

5.1 PARSEME framework

PARSEME initiative (Savary et al, 2017b), which was mentioned earlier, aims at the development of methods, tools and resources for multilingual MWE processing. The first phase of the project was devoted to VMWEs. A unified VMWEs typology and annotation guidelines have been developed and validated by multilingual pilot annotations. These guidelines were structured as decision trees over basic linguistic tests, so as to optimise the reproducibility of the annotator's decisions, which was later confirmed by the good inter-annotator agreement scores. Then, several annotation campaigns resulted in corpora covering 26 languages and containing millions of words and dozens of thousands of annotated VMWEs. These corpora were distributed under Creative Commons licenses (CC BY and CC BY-NC-SA[1]). The only exception to this rule was precisely an Arabic corpus, created by Hawwari et al. in PARSEME edition 1.1 (Savary et al, 2018a), which remained inaccessible. Consequently, we proceeded with the creation of the new corpus from scratch.

PARSEME's universality-driven effort still leaves room for language-specific phenomena that could add new decision trees or categories. Our previous work (Hadj Mohamed et al, 2022) revealed that the PARSEME guidelines apply to Arabic with only minor adjustments, and with no Arabic-specific features.

In addition to the guidelines and the annotated corpora, the PARSEME community has defined measures and constructed tools for evaluating both the quality of the annotation and VMWE identification systems[2].

5.2 Source data

As we wish our VMWE-annotations to be released openly, we chose the only Arabic corpus whose data are released under an open license, the Prague Arabic Dependency Treebank (PADT) (Hajic et al, 2009). It currently has 7,664 annotated sentences from various newswire sources[3], namely the Agence France Presse, the Al Hayat News Agency and the Xinhua news agency. PADT is part of Universal Dependencies (UD[4]), a universality-driven project for morphosyntactic annotation. PARSEME largely and increasingly relies on UD. In particular the CUPT data format used by PARSEME is an extension of the UD CoNLL-U format. Thus, adapting PADT to PARSEME is straightforward.

5.3 Annotation process

Based on the PADT data, we proceeded with the annotation of Arabic VMWEs using the PARSEME guidelines. Therein, a VMWE is identified on the basis

[1] https://creativecommons.org/licenses/by/4.0/
[2] https://gitlab.com/parseme/utilities
[3] In Arabic, the national newspapers are intended for all Arab countries. This means that whether the newspaper is Tunisian, Lebanese or other, it mostly uses MSA and delivers news on other Arab countries. This gives us an idea of the nature of the texts of the PADT.
[4] https://universaldependencies.org/

of its canonical form, however, in the corpus non-canonical forms (e.g. passive, or a clause with an extracted component) are also to be annotated.

Recall that the PARSEME guidelines are conceived as decision trees, so as to ensure reproducibility of the VMWE tagging and categorization. The decision trees incorporate both universal and language-specific tests. Universal tests take into account general criteria that are applicable to all languages, while language-specific tests focus on the structural, lexical, morphological, and syntactic characteristics of the individual languages.

Based on these guidelines, we manually annotated VMWE occurrences in PADT. 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, and the subsequent tests are applied to this form. Secondly, we determine the lexicalized components. Thirdly, we apply the generic decision tree[5], which – on the basis of the candidate's syntactic structure – redirects us towards decision subtrees specific to various VMWE categories. After these tests, we are able to decide whether the candidate is indeed a VMWE, and, if so, what is its category.

5.4 Corpus quality

Our previous work (Hadj Mohamed et al, 2022) involved creating a subset of 1,062 sentences from the PADT corpus, independently annotated by two native Arabic speakers. This initial phase revealed that inter-annotator agreement (IAA) is encouraging already at an early stage of the annotation process. This confirmed that the PARSEME guidelines are reasonably applicable to Arabic.

A_1	A_2	F_{span}	κ_{span}	κ_{cat}
763	704	0.699	0.626	0.864

Table 1 Inter-annotator agreement on a sample of 1,062 sentences, with A_1 and A_2 VMWEs annotated 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 1 shows the IAA calculated using PARSEME tools[6]. As discussed in Ramisch et al (2018), F_{span} is the MWE-based F-measure of A_1's annotations with respect to A_2, and vice versa. This measure defines an annotation as correct if both annotators identify identical tokens as being a part of the same VMWE (i.e., partial overlaps are totally incorrect). κ_{span} then calculates observed agreement corrected by expected agreement. The expected agreement can be calculated under the assumption that the total number of decisions which annotators have to take can be roughly estimated by the number of verbs in the annotated text (a VMWE usually contains a verbal head). For categorization, Cohen's kappa κ_{cat} is calculated on the VMWEs identified by the two annotators with the same components.

[5] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/
[6] https://gitlab.com/parseme/utilities

We could not afford a parallel annotation and adjudication of all the data, but to increase the quality of the final corpus, we enhanced our annotations with an additional step called *consistency check*, which is offered by the PARSEME tools[7]. This step handles inattention errors and inconsistent annotations by clustering all positive and negative examples of a specific VMWE encountered in the whole corpus.

To evaluate the impact of the consistency check's improvement, we retrained MTLB-STRUCT (cf. Section 8) on the data preceding the consistency check. This examination demonstrated an increase in the MWE-based[8] score from 56.27% to 60.49%.

Let us consider example (20). Fig. 1 shows the interface of the consistency checking tool, in which we notice that we have forgotten an idiomatic occurrence (underlined in red) of the expression from (20). We can rectify this by choosing the right type from the suggested list. The second case, underlined in blue, shows a fortuitous co-occurrence (example 21) of the same components of the expression that should not be corrected. Other cases of inconsistency may also occur, such as assigning the wrong type to an expression, forgetting to select all components of a certain expression, etc. All these errors can be manually corrected in the interface and the corrections are then pushed to the original annotated files.

(20) الأحكام طعن في
al-ahkām fī ṭ'n-
the-judgements in stab.HE.SG.PAST
lit. he stabbed in the judgements / 'he appealed the judgements'

(21) لأحكام وفقا الطعن
al-ahkām ūfqā ṭ'n
to-judgements according stabbing-the
lit. the stabbing according to judgements / 'the appeal according to judgements'

6 Corpus statistics

Table 2 presents the overall statistics of the previously unreleased PARSEME Arabic corpus by Hawwari et al. (cf. Section 5.1) and our newly created PARSEME-AR corpus (edition 1.3). In the Hawwari et al. corpus, 4,219 VMWEs were annotated across 3,137 sentences. These VMWEs divide into 42% LVCs.full, 31% VIDs, 26% VPCs.full, 0.4% IRVs, 0.8% MVC, 0% LVC.cause and 0% VPCs.semi. However, this corpus was never officially released.

[7] https://gitlab.com/parseme/corpora/-/wikis/annotating-new-corpora
[8] The MWE-based measure is the F1-score for fully predicted MWEs.

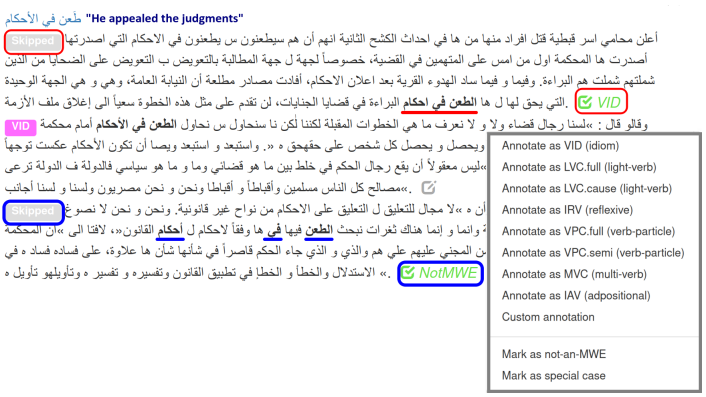

Fig. 1 Consistency check of the VMWE in example (20)

Our new corpus, named PARSEME-AR, contains a total of 4,749 VMWEs spread across 7,483 sentences. The most frequently occurring VMWE category was LVC.full, accounting for 56% of the total, followed by VID at 25%, and MVC was the least prevalent category, accounting for only 0.1% of the total. The quasi-universal IRV and VPC categories, which are widespread in Romance, Slavic, and German languages, but absent or uncommon in other languages, do not exist in Arabic. The density of VMWEs in our corpus is approximately 0.63 per sentence.

# Corpus	# Sent.	# Tok.	VMWE occurrences								
			VID	IRV	LVC.full	LVC.cause	VPC.full	VPC.semi	IAV	MVC	All
Hawwari et al.	3,137	265,244	1,320	17	1,769	0	1,080	0	0	33	4,219
PARSEME-AR	7,483	311,743	1,182	0	2,678	303	0	0	581	5	4,749

Table 2 Statistics of the Hawwari et al corpus and our 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.

This result contrasts with the statistics of the previous unreleased corpus, specifically in terms of the absence of VPCs. We claim that particles, as defined in PARSEME, are either non-existent or very rare in Arabic. In fact, the term "particle" is often used in MSA to denote various function words such as proclitics or prepositions. In the PARSEME guidelines, conversely, prepositions are disjoint from particles in that the former do and the latter do not take complements. Thus, the VPCs from the Hawwari et al. corpus may actually be IAVs.

The PARSEME shared task 1.2 focused on identifying unseen VMWEs, where a VMWE is considered unseen if its multiset of lemmas was never annotated as a VMWE in either the training or development data. The evaluation criteria include not only the overall F1 score but also the performance on the

unseen VMWEs. Thus, the corpora have been split so that the test set contains at least 300 unseen VMWEs per language. According to the split method described by Ramisch et al (2020), our corpus, used for the MWE identification task (cf. Section 8), was split into training (TRAIN), development (DEV), and test sets (TEST). Table 3 displays the statistics of the resulting splits for the corpus. The splitting was carried out with a target number of 300 unseen VMWEs (in the test set) and 10 random splits.

PARSEME-AR in its current state is already available under the PARSEME repository[9] under the CC-BY v4 license[10], including the double-annotated IAA sample and the split of the corpus. Thus, the results presented here are fully reproducible, using the PARSEME utilities[11]. A brief description of the files is also provided in the README under the PARSEME repository. Additionally, an immutable version of the corpus is part of the PARSEME corpus release 1.3.[12]

	# Sent.	Unseen	LVC.full	LVC.cause	VID	IAV	MVC	Total
TRAIN	6091	-	2178	236	955	468	4	3841
DEV	342	100	121	15	54	38	0	228
TEST	1050	300	379	52	173	75	1	680

Table 3 Statistics of the annotated corpus divided into TRAIN, DEV and TEST: number of sentences, total number of VMWEs (Col. 9), as well as a breakdown by category of VMWE. (Col. 3) provides statistics for the number of unseen expressions, i.e. those which occur in DEV and TEST but do not occur in TRAIN and those which occur in the TEST and not occur in TRAIN+DEV, respectively.

7 Linguistic observations

In this section, we will highlight some linguistic characteristics of the VMWEs present in our corpus. It is important to mention that, even though Arabic has many dialects, no dialect-specific VMWEs were found in our data, which can be attributed to the nature of the source data. Nonetheless, the VMWEs do exhibit properties or tendencies that are specific to Arabic compared to other languages. We will focus on these properties in this section.

[9] https://gitlab.com/parseme/parseme_corpus_ar
[10] https://creativecommons.org/licenses/by/4.0/
[11] https://gitlab.com/parseme/utilities/-/blob/master/st-organizers/corpus-statistics/mwe-stats-simple.py and https://gitlab.com/parseme/utilities/-/blob/master/st-organizers/corpus-statistics/mwe-stats.py
[12] http://hdl.handle.net/11372/LRT-5124

7.1 Masdar

In MSA, Al-maSdar (literally 'source') is a linguistic term that refers to a nominal derived from a verb. It is used to express the action or state in an abstract and indeterminate way. In other words, the transformation from the verbal form to the nominal form through Al-maSdar does not alter the meaning of the utterance. This is achieved by creating a close relationship of synonymy, yet without any notable aspectual-temporal markings. In addition, Al-maSdar forms do not appropriate one grammatical or morphological category. They can be nouns, sometimes verbal nouns (which are similar to gerunds in English), and sometimes nominalizations.

In compliance with the PARSEME guidelines, annotations should be made for morphosyntactic variants of a VMWE that retain their original meaning. Thus, if a verb or noun in a VMWE is transformed into its Al-maSdar form while preserving the idiomatic meaning, it is considered a valid occurrence of a VMWE. For instance, in example (22), Al-maSdar تقديم ('giving') derived from the verb قدّم ('give') behaves as a light verb and the meaning is carried by the noun نصح ('advice'). In this case, the candidate VMWE is النصح تقديم (lit. 'giving advising') 'the giving of advising', and the canonical form to which the linguistic tests are applied is (23), which passes the LVC.full tests.

(22) النصح تقديم اسطعت كلما عليك
al-nṣḥ **tqdym** iastṭ't klmā 'līk
the-advising give.MSDR you.can whenever on-you
lit. giving the advising whenever you can is on you / 'you should give advice whenever you can'

(23) النصح تقدّم أن اسطعت كلما عليك
al-nṣḥ **tqdm** an iastṭ't klmā 'līk
advise.MSDR give.2.SG.PRES to you.can whenever on-you
lit. that you give advising whenever you can is on you / 'you should give advice whenever you can'

In an LVC, Al-maSdar can take the role not only of the light verb but also of the predicative noun. An interesting specificity of Arabic is when the verb

and Al-maSdar are derived from the same verbal root, which leads to semantic duplication, we note that in example (24) the noun قتالا ('fighting') can be actualized by the verb قاتلت ('she fought'), both are from the same root. In other words, in this expression, the same verb plays the role of a (semantically void) light verb and of the (semantically rich) predicate describing the action. This relates to the phenomenon of lexical couplets, typical for Arabic. Such *verb/Al-maSdar* combinations pass the LVC.full tests.

(24) قاتلت قتالا
qtālā qātl-t
fight fight.FEM.SG.PAST
lit. she fought fighting / 'she fought'

Al-maSdar can also occur in VIDs, like in example (25), Al-maSdar تبيض ('whitening') is the nominal variant of the verb بيض ('whiten') and the idiomatic meaning is kept in this sense so we annotate it as VID.

(25) أموال تبيض
amwāl tbīḍ
whiten.MSDR money
lit. money whitening / 'money laundering'

7.2 Discontinuities

Since the order of words is relatively free in MSA, we can have many insertions between VMWE components, which makes the VMWE hard to handle for both humans and machines. A frequent source of discontinuities are grammatical words like determiners, auxiliaries, negations, etc. A VMWE may have only one inserted word as in example (4), where the light verb and the predicative noun are separated by the definite article ال 'the' (attached at the front of the noun but considered a separate token). However, sequences of complex arguments and adjuncts can be inserted between a verb and an object, as in example (26). It contains 9 insertions between the VMWE components إتصال تلقى ('receive a call'). This phenomenon is potentially frequent due to the standard VSO order of Arabic.

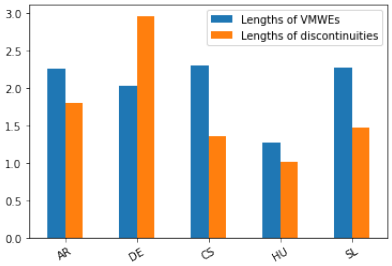

Fig. 2 Ranking of Arabic (AR), Czech (CS), German (DE), Hungarian (HU) and Slovene (SL) in terms of length and discontinuities of VMWEs

(26) هاتفيا اتصالا الاحد اليوم صباح هنا ماهر احمد
hātfīā **atṣālā** al-aḥd al-īūm ṣabah hnā Maher Ahmed
phone call the-Sunday the-today morning here Maher Ahmed
المصرى الخارجية وزير تلقى
al-mṣri al-ẋārğīT ūzīr **tlqi**
the-Egyptian the-foreign minister receive.3.SG.PAST
lit. the Egyptian Foreign Minister Ahmed Maher received a phone call here on Sunday evening

Lengths of VMWEs					Lengths of discontinuities					
Avg	%1	%2	%3	%>3	Avg	%0	%1	%2	%3	%>3
2.25	0.36	77.34	19.92	2.38	1.79	43.15	24.57	10.78	7.18	14.32

Table 4 Lengths and discontinuities of the Arabic VMWE occurrences: average number of tokens (Col. 1); percentage of VMWEs with 1, 2, 3 and more than 3 tokens (Col. 2–5); average length of discontinuities (Col. 6); percentage of VMWEs with discontinuities of length 0, 1, 2, 3 and more than 3 (Col. 7–11).

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. Following Savary et al (2018b), we analysed the corpus in terms of lengths (numbers of tokens) of the annotated VMWEs and of their discontinuities (numbers of external tokens occurring between the first and the last token of a VMWE). Table 4 shows the results of this analysis. In particular, over 77% of all VMWEs contain 2 tokens (column 3), above 43% are continuous (column 7) but more than 14% (last column) have more than 3 gaps.

We compare these results to the 18 languages from the PARSEME suite in edition 1.0.[13] as shown in Figure 2. The average length of Arabic VMWEs (2.25 in column 1 of Table 4) is not an outlier, since in 17 (out of the 18) languages

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

this factor is between 2.02 and 2.71. The non-existence or rareness of single-token VMWEs[14] (0 in column 2 of Table 4) is also a feature of 14 languages in the PARSEME suite (Hungarian, German and Portuguese being outliers in this category). However, Figure 2 shows that, in terms of discontinuities, Arabic is the second most outstanding language (after German). It has 1.79 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 VMWEs (43.15%) and the 2nd highest percentage of VMWEs with gaps longer than 3 (14.33%), after German (35.7% and 30.5%, respectively). This can be due, inter alia, to two combined phenomena: the dominance of the VSO sentence order in Arabic, and the fact that many VMWEs consist of a verb and its direct object, as illustrated above with example (26).

7.3 Word order and pronominalization

An interesting interdependence between the word order and verbal morphology in Arabic is illustrated in example (27). It contains an LVC.full in which the predicative noun نظر وجهة (**nżr ūğhT** / lit. 'view destination') 'point of view' is a (nominal) MWE itself and is the direct object of the light verb طرح 'throw'. Recall that the standard word order in Arabic is VSO. Here, however the sentence starts with the object and ends with the verb (the subject is dropped). When this situation occurs in Arabic, the verb systematically bears an anaphoric pronoun which refers to the object. Here, the verb طرح (ṫrḥ) 'asks' is agglutinated with the anaphoric pronoun ها (hā) 'her' referring to the nominal head وجهة (ūğhT) 'destination'. Similar constructions occur in verb-object VMWEs of type VID.

(27) طرحها أخرى نظر وجهة
ṭrḥ-hā aẋri **nżr** **ūğhT**
throws-her.3.SG.FEM another view destination.SG.FEM
lit. asks another view destination / 'puts forward another point of view'

This systematic anaphora conditioned by the word order sheds new light on the phenomena on the frontier between MWEs and coreference. Past studies, e.g. (Moon, 1998; Savary et al, 2023), hypothesise that, due to semantic non-compositionality of MWEs, proper subsets of MWEs are normally not subject

[14] Note a token can contain several agglutinated words, so it can, indeed, be a MWEs.

to pronominalization and, more generally, are not expected to occur in co-reference chains. It seems that this statement should be mitigated for Arabic, where pronominalization of a nominal group, here نظر وجهة (lit. 'destination view') 'point of view', in an object-initial sentence is compulsory, whether the verbal phrase is semantically compositional or not.

7.4 Lexicalized preposition in LVC

Another interesting observation is the required presence of a specific preposition to have a light verb. An Arabic distributional verb[15] can be transformed into a light verb by using a preposition. Adding this preposition completely changes the meaning of the verb to which it is applied. This phenomenon is common in MSA. For instance, the verb قـام ('to stand up') loses its distributional value and becomes a light verb after the addition of the preposition ب ('with'). This shift from ordinary verbal value to light value is illustrated by the examples (28) and (29).

(28) قام من دون تفكير إلى صلاة الجماعة
qām mn dūn tfkīr ili ṣlāT al-ğmā'T
stand.up.3.SG.PAST from without thinking to prayer the-group
lit. he stood up without thinking to the congregational prayer.

(29) قام الوزير بزيارة موريتانيا
qām al-ūzīr **b-zīārT** mūrītānīā
stand.up3.SG.PAST the-minister with-visit Mauritania
lit. 'the minister stood up with a visit to Mauritania' 'the minister payed a visit to Mauritania'

We notice that in the first example (28), the verb قام 'stand up' fulfills its usual function as a full-meaning verb and that the preposition إلى 'to' is not required by the verb but introduces a circumstantial complement. Conversely, in (29) the preposition ب 'with' is required and makes the verb lose its literal sense. The meaning here is carried by the predicative noun زيارة 'visit' whereas the verb قام 'stand up' behaves as a light verb.

[15] a verb which in each context has a particular meaning

8 Benchmark for MWE identification

MWE identification is a complex task due to the complex linguistic nature of MWEs. In this section, we present this MWE-related task on MSA data. The purpose of automatic MWE identification task is to label the words that make up MWEs. We propose to analyze the robustness of two systems for identifying MWE in MSA namely Seen2Seen (Pasquer et al, 2020) and MTLB-STRUCT (Taslimipoor et al, 2020). The two systems are briefly discussed, and their performance in Arabic is compared to that of other languages. These results are significant, since they constitute the first benchmark in this field for Arabic and show that our corpus is exploitable.

8.1 Training state-of-the-art systems on Arabic

The first system, Seen2Seen[16] created by Pasquer et al (2020), ranked first in the global F-measure for the closed track[17]. Seen2Seen reads all of the annotated VMWEs in the training corpus and then extracts all candidate occurrences of the same set of lemmas from the test corpus. These candidates are then passed through a series of morpho-syntactic filters. A total of 8 filters are defined, and during the training phase we can decide which filter to activate for which language based on the performance on the development corpus. Filter 1 (f1) ensures disambiguation of components[18]. Filters 2 and 3 (f2 and f3) verify the specific order of components within VMWEs, considering both ordered POS sequences (f2) and discontinuities (f3) within VMWEs. Filter 4 (f4) excludes candidates whose discontinuity length is higher than the highest length observed in training for the given category. Filter 5 (f5) enhances f4 by prioritizing candidates with the lowest discontinuity length. Filter 6 (f6) focuses on syntactic connectivity, retaining candidates which form connected dependency subtrees or which are connected by a syntactic chain with limited insertion. Filter 7 (f7) emphasizes nominal inflectional inflexibility, retaining candidates with nominal components matching the inflection of seen VMWEs. Filter 8 (f8) verifies if all occurrences of the candidate VMWE have the same nested annotation, whatever was the context. Thus, f8 imitates the annotation in the training corpus to see whether to keep embedded candidates.

We used the split corpus described in Table (3) as our input to both systems. For Seen2Seen, the best system results for Arabic are displayed when f4 and f8 are activated, while the other filters are not. For instance, these two

[16] https://gitlab.com/cpasquer/st_2020

[17] In the closed track systems used only the provided training and development corpora (with VMWE and morpho-syntactic annotations) and the provided raw corpora.

[18] Lemmas which may be shared by different words are disambiguated based on their part of speech (POS).

filters accept the two VMWE candidates included in example (30). Here, the VMWE دراسة يتم اجراها 'study making its procedure' meaning 'to proceed to a study' contains another VMWE يتم اجراها 'making its procedure' meaning 'to proceed'. However, a comparable F-score can be achieved by activating filters 1 and 7 as well. Filters 2 and 3 might be omitted from activation since, as previously noted, Arabic allows for a flexible word order, and their activation could potentially exclude suitable candidates. Filter 6 may not be activated either, as Arabic exhibits a high degree of discontinuity, resulting in many insertions that do not align well with the filter's criteria, which favors connected candidates with limited insertions.

(30) اجراها يتم دراسة هناك
ağrāhā **ītm** **drāst** hnāk
proceeding.it making.it study there.is
lit. there is a study making its proceeding / 'there is a study in progress'

The second system, MTLB-STRUCT, created by Taslimipoor et al (2020) is built upon a BERT model and fine-tuned using a multi-task approach for parsing and MWE identification. The weights of the BERT model are shared by the two tasks. A fully connected layer that performs sequence labeling is added as the final layer for the first task, MWE tagging. For the second task, linear layers and dependency CRF modules are developed simultaneously. The system participated in the 2020 PARSEME Shared Task open track[19] and ranked first according to the average F1-score across all 14 languages, both for unseen and seen VMWEs.

We tested the model in the two settings: the performance on the single task (the model is back-propagated based on the MWE-specific loss function), and the performance on the multi task (the model is based on the multi-task loss function).We used the same default settings of the pre-trained model as for all the other languages, which is bert-base-multilingual-cased. We trained the models for 10 epochs. The maximum lengths of sentences was 250 for training which was chosen based on the word piece tokenisation of multilingual BERT.

8.2 System results

We evaluate MWE identification systems using standard measures, which have been applied for PARSEME evaluation campaigns (Savary et al, 2017b). We used the evaluation metrics in accordance with the shared task criteria: MWE-based[20] precision (P), recall (R) and F1 measures for all VMWEs and unseen

[19] In the open track systems may use the provided training corpora, plus any additional resources deemed useful (MWE lexicons, symbolic grammars, wordnets, other raw corpora, word embeddings, language models trained on external data, etc.).

[20] accurately detecting the whole VMWE

ones (Unseen MWE-based[21]). We also consider that efficiency in identifying the individual components of a VMWE (token-based[22]

P, R, and F1) is useful to be reported.

Table 5 summarises the results of these metrics on our Arabic corpus and four additional languages (Arabic AR, German DE, Hebrew HE, Polish PL and Italian IT) from the PARSEME shared task. The discontinuity ranking, the number of annotated VMWEs, and the sizes of the corpora all play a role in our selection of these languages. Since Arabic and Hebrew are both Semitic languages with close corpus sizes, comparing them should be meaningful. Italian is retained since it has a similar number of annotated VMWEs (4749 and 4257 in Arabic and Italian, respectively). Comparing Arabic to German is interesting due to both languages being outliers as far as discontinuity of VMWEs is concerned (cf. section 7.2). Finally, Polish was added because both systems achieved the highest scores in the PARSEME shared task for this language. Thus, it may serve as an upper bound for the state-of-the-art performances.

As seen in Table 5, typically, the precision values are higher than recall. This can be justified by the fact that, as noticed in the PARSEME shared task edition 1.1, identifying unseen VMWEs is particularly hard. For this reason, edition 1.2 focused on unseen VMWE identification, and the corpus splits were performed so as to keep at least 300 unseen VMWEs in each test corpus. We follow the same split method for Arabic (cf. section 6), which results in 46% of all VMWEs in test being unseen in TRAIN+DEV. Understandably, in Arabic, like in most languages, for both systems MWE-based scores are lower than their token-based scores.

Based on the observed results MTLB-STRUCT outperforms Seen2Seen for German, Hebrew and Arabic with MWE-based F1-scores of 76.17%, 48.3% and 60.49%. Conversely, Seen2Seen F1-scores are slightly higher for Italian and Polish (64.92% and 81.85%, respectively). Since Seen2Seen was designed to only capture seen VMWEs, MTLB-STRUCT is much more effective at capturing unseen VMWEs for all languages. Still, its unseen-based F1-scores for Hebrew and Italian are rather low, not exceeding 19.59% and 20.81%, respectively.

Among the five languages, systems show more modest performance for Semitic languages for both MWE-based F1-scores and unseen-based F1-scores. This can be explained by: (i) errors in the morphological/syntactic annotation, which was performed automatically, (ii) the morphologically rich nature of these languages.[23] The issues with morphosyntax annotation errors in PADT are discussed in more detail in Hadj Mohamed et al (2022).

An interesting opposition for MTLB-STRUCT appears when comparing the global scores with those for continuous and discontinuous VMWEs, as shown in Table 7. Usually, F1-scores for continuous VMWEs are largely higher than for discontinuous ones (Markantonatou et al, 2018), whereas the opposite proves true for Arabic. Namely, MTLB-STRUCT scores by over 2 points higher for discontinuous VMWEs than for continuous ones (59.70% and 57.60%,

[21] accurately detecting the VMWEs unseen in TRAIN and DEV

[22] The token-based measure is the F1-score allowing for partial matches of VMWE components.

[23] While morphosyntactic annotation of the Italian and German corpus was also performed automatically, these languages are less morphologically rich. Polish is morphologically rich but has mostly manual annotations in the morphosyntactic layers.

respectively). This results are probably correlated with the density of discontinuous VMWEs that is very high (around 57% of VMWEs are discontinuous) (cf. Section 7.2). We also observed a notable discrepancy in the performance between Arabic and Hebrew as well as Italian for unseen MWEs using the MTLB-STRUCT approach, as shown in Table 5. A possible explanation for this difference could be that unseen expressions in Arabic might share more similarities with seen expressions in the training data compared to Hebrew and Italian. For instance, the unseen VMWE إتخذ قرار 'make a decision' was easily identified in the test data because it closely resembles the VMWE أخذ قرار 'make a decision' seen during training. Both expressions share the same core noun قرار (decision) and verbs with similar meanings (إتخذ and أخذ) (take and make). This overlap in lexical components and syntactic structure likely facilitated the model's ability to generalize the new expression accurately.

In Table 6, the results of the MWE-based metrics can be compared per category: LVC, VID, IAV and MVC. These results show that MTLB-STRUCT performs better for LVC.full and IAV with F1-score of 57.72% and 69.94%. This might be related to the fact that LVCs are more frequent and follow relatively productive patterns (e.g. they use frequent light verbs, their predicative nouns have common semantic properties), which might be relatively easy to generalize in neural networks. Seen2Seen, conversely, performs better for VIDs (and outperforms MTLB-STRUCT by 4.86%). The precise reason of this fact calls for more in-depth analyses.

8.3 Error analysis

For Seen2Seen, errors are due to false negatives which affect the recall. False negatives are due to : (1) unseen VMWEs, which, by the definition of the task, are not identified, (2) false tokenization as in (31): the first sequence ولعبوا 'and they.played' should be tokenized into two words و 'and' and exidioلعبوا they.played, and (3) annotation problems due to vowelization. In fact the majority of texts in Arabic are written using non-vowelized letters, but in some cases we have encountered different lemmas for the same word, in which letters were sometimes voweled and sometimes not. Seen2Seen obviously considers such cases as different lemmas, and fails to extract differently spelled multi-sets of lemmas as VMWE candidates.

Lang			**AR**	DE	HE	PL	IT
MTLB-STRUCT	MWE-Based	P	61.47	77.11	56.2	82.94	67.68
		R	59.54	75.24	42.35	79.18	60.27
		F1	**60.49**	76.17	48.3	81.02	63.76
	Token-Based	P	69.97	83.18	68.37	85.06	74.46
		R	67.31	74.75	44.82	79.42	62.08
		F1	**68.61**	78.74	52.99	82.14	67.71
	Unseen MWEs	P	39.00	49.17	25.53	38.46	20.32
		R	36.00	49.5	15.89	41.53	21.33
		F1	**37.44**	49.34	19.59	39.94	20.81
Seen2Seen	MWE-Based	P	58.33	86.21	65.84	91.15	67.76
		R	45.29	57.65	31.81	74.28	62.31
		F1	**50.99**	69.09	42.9	81.85	64.92
	Token-Based	P	63.29	89.07	68.37	91.74	71.76
		R	45.82	52.18	30.93	73.38	60.27
		F1	**53.15**	65.8	42.6	81.54	65.52
	Unseen MWEs	P	0	12.5	0	100	20
		R	0	0.33	0	0.33	0.33
		F1	0	0.65	0	0.66	0.66

Table 5 Result obtained by Seen2Seen and MTLB-STRUCT: The MWE-based, token-based and unseen-based metrics for Arabic, German, Hebrew, Polish and Italian

(31) دورا و-لعبوا
dawran la'ibo-u
role and-play.3.PL.PAST
lit. and they played a role

(32) سؤال طرَح
suāl ṭraḥ
question ask.3.SG.PAST
lit. he asked a question

System	LVC.full			LVC.cause			VID			IAV			MVC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Seen2Seen	61.81	41.42	49.61	63.16	23.08	33.80	71.43	43.35	53.96	32.67	67.12	43.95	0	0	0
MTLB-STRUCT	55.47	60.16	57.72	55.00	42.31	47.83	50.93	47.40	49.10	63.33	78.08	69.94	0	0	0

Table 6 The results of MWE-based F-measure for Seen2Seen and MTLB-Struct evaluation per MWE category.

System	Continuous MWE-based			Discontinuous MWE-based		
	P	R	F1	P	R	F1
Seen2Seen	60.48	51.79	55.80	54.64	36.55	43.80
MTLB-STRUCT	58.06	57.14	57.60	60.61	58.82	59.70

Table 7 MWE-based F1-measure for Seen2Seen and MTLB-STRUCT in terms of continuity and discontinuity.

(33) سؤال طرح
suāl ṭrḥ
question ask.3.SG.PAST
lit. he asked a question

For instance, example (33) occurring in TEST is the same VMWE as in (32) occurring in TRAIN but the former was missed by Seen2Seen, due to the non-vowelled spelling of the verb طرح 'ask'.

Other false negatives are due to the selection of filters: filter 4 excludes the discontinuous candidates in which the distance between different components is higher than encountered in TRAIN for the same VMWE category.

As far as false positives of MTLB-STRUCT are concerned, we observe verb-noun combinations in which the verb is a frequent light verb. For instance, 22% of all VMWEs in TRAIN contain one of the two verbs قام 'stand-up' or يملك 'have' as in example (29). The same verbs are also frequent outside of VMWEs, and their combinations with nouns are frequently mistaken for VMWEs by MTLB-STRUCT, as is (34).

(34) الانتخابات في المشاركة من الاخوان منع
al-āntxābāt fī al-mšārkT mn al-āxwān mn'
the-elections in the-participation from the-brotherhood preventing
يقوم على
'li īqūm
on stand-up.3.SG.PRESENT
lit. 'it stands up on preventing the Brotherhood from participating in the elections ' 'It is based on preventing the Brotherhood from participating in the elections'

Besides, both systems mistake some literal occurrences for true VMWEs. For instance, the VMWE in (35) occurs in TRAIN. Example (36) contains the same lemmas but in a non-idiomatic combination. Seen2Seen as MTLB-STRUCT, however were unable to filter out this candidate.

(35) الرف وضعه على يتم الاتفاق
al-rf **'li** **ūḍ'-h** ītm al-ātfāq
the-shelf on put.IT.PRESENT complete.IT.PRESENT the-agreement
لتنفيذ بالنسبة أما
l-tnfīḍ b-āl-nsbT amā
for-doing for-about as.for
lit. as for doing the agreement, complete put it on the shelf / 'as for the implementation of the agreement, it is put in the state of disuse'

(36) الرف على وضعه أستطيع الباقي (*NOT-MWE*)
al-rf 'li ūḍ'-h asttṭī' al-bāqī
the-shelf on put.it can.1.SG.PRESENT the-rest
lit. the rest I can put it on the shelf

Also, false negative are produced by MTLB-STRUCT due to the high discontinuity of some expressions. For instance, in example (37) the system succeeds to identify the first occurrence of the VMWE عقد إجتماع (lit. 'tie a meeting') 'hold a meeting' when its components are close (separated by one element) whereas it fails to identify its second occurrence in the same sentence when we have insertions.

(37) التسعة لمنصب الرئيس الليلة اول إجتماع
iğtmā' aūl al-līlT al-r'īs l-mnṣb al-ts'T
meeting first the-tonight the-president for-position the-nine
المتحدة ... ... يعقد الديمقراطيون المتنافسون
al-mtnāfsūn al-dīmqrāṭīūn **ī'qd** ··· al-mtḥdT
the-contenders the-democratic tie.THEY.PRESENT ··· the-united
المرشحين الديمقراطيين للرئاسة فى الولايات
al-ūlāīāt fi l-al-r'iāsT al-dīmqrāṭyin al-mršḥīn
the-state in for-the-presidential the-democratic the-candidates
عقد اول إجتماع بين
bīn **iğtmā'** aūl **'qd**
between meeting first tying
lit. 'tying meeting between the candidates the democratic for the presidential in the united state ... they tie the democratic contenders the nine for the position the president tonight first meeting ' / 'The first meeting was held between the Democratic presidential candidates in the United States... The nine Democrats competing for the position of president are holding their first meeting tonight'

9 Related works

Several studies have focused on building monolingual corpora annotated with VMWEs. Notable examples include corpora developed for the PARSEME shared task (Savary et al, 2017a; Ramisch et al, 2018). Edition 1.2 of this task covered 14 languages, including non-European languages such as Chinese, Hindi, and Turkish. Edition 1.3 of the PARSEME corpus, released independently of any shared task, gathered data released previously in editions 1.0, 1.2 and 1.3, added new languages, increased or enhanced some annotations, and made them fully compatible with Universal Dependencies. This effort resulted in a unified corpus in 26 languages, including our own Arabic corpus described in this paper. Vincze and Csirik (2010) manually annotated a Hungarian corpus with light verb constructions, highlighting their utility for tasks like machine translation and information extraction. In the domain of monolingual English corpora, significant contributions include the "MWE-aware English Dependency Corpus" provided by the Linguistic Data Consortium (Kato et al, 2016). This corpus, focusing on compound words, is used for training parsing models. Vincze et al (2011) introduced Wiki50, which comprises 50 Wikipedia articles (equating to 4,350 sentences) annotated with various MWE types, including compounds, verb-particle constructions, idioms, light verb constructions, multi-word verbs, and named entities.

Although MWEs in Arabic have been found to be frequent and particularly challenging for learners, usage-based accounts of these structures in other

languages like in English by far outnumber those for Arabic. Indeed, a comprehensive annotation and corpus-based analysis of MWEs in MSA have not yet been attempted despite the potential benefits such a resource and analysis could provide for linguists, translators, language instructors, and learners alike. However several studies have been carried out on Arabic MWEs (AMWEs).

Attia (2006) performs a MWE-aware parsing of Arabic texts with finite-state machinery and Lexical Functional Grammar (LFG). Later, Attia et al (2010) implement a semi-automatic linguistic method based on regular expressions for extracting MWEs in Arabic texts. They propose 3 approaches that focus on nominal AMWEs. The first approach finds correspondence asymmetries between titles of Wikipedia pages in Arabic and in 21 other 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) creates a list of MWEs in the Egyptian dialect based on a collection of 5,000 expressions manually extracted from Arabic dictionaries and grouped into syntactic types. Their final list comprises 4,209 MWEs: Verb-Verb Construction (VVC), Verb Noun Construction (VNC), Verb-Particle Construction (VPC), Noun-Noun Construction (NNC) and Adjective Noun Construction (ANC). After that they run a pattern-matching algorithm on a large part of the Arabic Gigaword and they find 481,131 MWE instances for 250 million tokens.

Abdou (2012) explores an 83-million-word Arabic corpus in order to examine AMWEs, mainly MSA idioms, with regard to their semantic, discursive, lexical and grammatical properties. He establishes the main patterns of the linguistic behavior of AMWEs and develops an empirical taxonomy of six AMWE types: verb-subject, verbal, nominal, prepositional, adjectival and adverbial idioms. A example of the first type is أفل نّجم (**afala najmu** / lit. 'the star set') 'the glory or fame of somebody/something ended'. The second type gathers verb-object combinations like in example (13). The author stresses the pervasiveness and lexical variability of these two types.

Ghoneim and Diab (2013) use the LDC GALE newswire parallel Arabic-English corpus to represent MWEs in a Statistical Machine Translation (SMT) pipeline. Various types of MWEs are considered: VMWEs (verb-noun constructions, verb-particle constructions, light verb constructions), noun-based MWEs

(noun-noun constructions, named entity constructions), adjective- and adverb-based MWEs. A list of MWEs extracted from English WordNet database 3.0 is also used and named entities are considered as a subtype of MWEs.

Hawwari et al (2014) describe an unsupervised approach to build a lexicon of Egyptian Multiword Expressions and a repository for their variation patterns. The lexicon contains 10,664 entries of Egyptian MWEs and collocations, linked to the repository.

Al-Badrashiny et al (2016) use a paradigm detector on the Arabic Treebank (ATB) (Maamouri and Bies, 2010) and Arabic Gigawords corpus to build a AMWEs resource. They manage to extract automatically 1,884 AMWEs. Each type of these 1,884 AMWEs has 20 different forms on average due to the morphological or inflectional changes of the AMWE components.

Zaghouani et al (2010) revise the original 493 Frame Files from the Pilot Arabic PopBANK and add 1462 new files for a total of 1955 Frame Files with 2446 frame sets including predicates such as light verb constructions and multi-word expressions.

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 5.1). 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.

10 Conclusion

In this paper, we introduce PARSEME-AR, a manually annotated Arabic VMWE corpus based on PARSEME guidelines. The corpus contains 4,749 annotated VMWEs divided into 56.39% of LVC.full, 24.88% of VID, 12.02% of IAV, 6.38% of LVC.cause and 0.10% of MVC. We observed a high rate of discontinuous expressions (58%) which can be explained by the word order which is relatively free and dominated by the VSO pattern in Arabic. We also established a state-of-the-art for the Arabic VMWE identification task by training and evaluating two of the best MWE identification systems from edition 1.2 of the PARSEME shared task, namely Seen2Seen and MTLB-STRUCT, on our data. MTLB-STRUCT outperforms SeenSeen with MWE-based F1 measure of 60.49% and 50.99%, respectively. We analyzed the systems' errors which provided hints towards possible enhancements in

methods for identifying VMWEs in Arabic.

In near future, we plan to make a slight adjustment to the guidelines that takes into account the annotation of anaphora. In addition, we aim to improve the data by resolving the issue of vocalization, modifying Seen2Seen so that it can handle different variants of a lemma, regardless of their vocalization. Furthermore, the extent to which resources developed for MSA can be applied to dialectal data is an intriguing prospect. Arabic dialects pose challenges in natural language processing due to their variability and limited resources compared to MSA. To effectively adapt MSA methods and tools to Arabic dialects, specific data collection, a deep understanding of dialectal linguistic characteristics, and the development of tailored techniques are essential. We consider this work an initial contribution for elaborating an Arabic universal terminology of VMWEs, which could ease the challenge of automatic processing of MWEs, in particular verbal ones.

References

Abdou A (2012) Arabic Idioms: A Corpus Based Study. Routledge, URL https://books.google.fr/books?id=nyk6bwAACAAJ

Al-Badrashiny M, Hawwari A, Ghoneim M, et al (2016) SAMER: a semi-automatically created lexical resource for Arabic verbal multiword expressions tokens paradigm and their morphosyntactic features. In: Proceedings of the 12th Workshop on Asian Language Resources (ALR12), pp 113–122

Attia M, Toral A, Tounsi L, et al (2010) Automatic extraction of Arabic multiword expressions. In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, pp 19–27

Attia MA (2006) Accommodating Multiword Expressions in an Arabic LFG grammar. In: International Conference on Natural Language Processing (in Finland), Springer, pp 87–98

Baldwin T, Kim SN (2010) Multiword Expressions. Handbook of Natural Language Processing 2:267–292

Benmamoun E (1997) Licensing of negative polarity items in Moroccan Arabic. Natural Language & Linguistic Theory 15(2):263–287

Boudchiche M, Mazroui A (2015) Evaluation of the ambiguity caused by the absence of diacritical marks in Arabic texts: statistical study. In: 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA), IEEE, pp 1–6

Constant M, Eryiğit G, Monti J, et al (2017) Survey: Multiword Expression processing: A Survey. Computational Linguistics 43(4):837–892. https://doi.org/10.1162/COLI_a_00302, URL https://aclanthology.org/J17-4005

Deksne D, Skadiņš R, Skadiņa I (2008) Dictionary of Multiword Expressions for translation into highly inflected languages. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)

Elkateb S, Black WJ, Rodríguez H, et al (2006) Building a Wordnet for Arabic. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)

Farghaly A, Senellart J (2003) Inductive coding of the Arabic lexicon. In: Workshop on Machine Translation for Semitic Languages: issues and approaches, New Orleans, USA, URL https://aclanthology.org/2003.mtsummit-semit.6

Ghoneim M, Diab M (2013) Multiword Expressions in the Context of Statistical Machine Translation. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, Nagoya, Japan, pp 1181–1187, URL https://aclanthology.org/I13-1168

Gross M (1982) Une classification des Phrases Figées du Français. Revue Québécoise de Linguistique 11(2):151–185

Hadj Mohamed N, Khelil CB, Savary A, et al (2022) Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure. In: 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp 1839–1848, URL https://hal.archives-ouvertes.fr/hal-03712937

Hajic J, Smrz O, Zemánek P, et al (2009) Prague Arabic dependency treebank: Development in data and tools. In: Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools

Hawwari A, Attia M, Diab M (2014) A framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). Association for Computational Linguistics, Doha, Qatar, pp 48–56, https://doi.org/10.3115/v1/W14-3606, URL https://aclanthology.org/W14-3606

Hawwari A, Bar K, 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, pp

24–29

Jackendoff R (1997) The Architecture of the Language Faculty. 28, MIT Press

Kato A, Shindo H, Matsumoto Y (2016) Construction of an English Dependency Corpus incorporating Compound Function Words. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp 1667–1671, URL https://aclanthology.org/L16-1263

Lyovin A, et al (1997) An Introduction to the Languages of the World. Oxford University Press, USA

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

Markantonatou S, Ramisch C, Savary A, et al (2018) Multiword Expressions at length and in depth: Extended papers from the MWE 2017 workshop. Language Science Press

Moon R (1998) Fixed expressions and idioms in English. Oxford University Press, Oxford

Nivre J, Hall J, Nilsson J (2004) Memory-based dependency parsing. In: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, pp 49–56

Pasquer C, Savary A, Ramisch C, et al (2020) Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut? In: Proceedings of the 28th International Conference on Computational Linguistics, pp 3333–3345

Ramisch C, Cordeiro SR, Savary A, et al (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). Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 222–240, URL https://aclanthology.org/W18-4925

Ramisch C, Savary A, Guillaume B, et al (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. Association for Computational Linguistics, pp 107–118, URL https://aclanthology.org/2020.mwe-1.14

SanJuan E, Ibekwe-SanJuan F (2006) Text mining without document context. Information Processing & Management 42(6):1532–1552

Savary A, Ramisch C, Cordeiro S, et al (2017a) The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, Spain, pp 31–47, https://doi.org/10.18653/v1/W17-1704, URL https://www.aclweb.org/anthology/W17-1704

Savary A, Ramisch C, Cordeiro SR, et al (2017b) The parseme Shared Task on Automatic Identification of Verbal Multiword Expressions. In: The 13th Workshop on Multiword Expression at EACL, pp 31–47

Savary A, Candito M, Mititelu VB, et al (2018a) Parseme multilingual corpus of verbal multiword expressions. In: Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop

Savary A, Candito M, Mititelu VB, et al (2018b) PARSEME Multilingual Corpus of Verbal Multiword expressions. In: Markantonatou S, Ramisch C, Savary A, et al (eds) Multiword Expressions at length and in depth: Extended papers from the MWE 2017 workshop. Language Science Press., Berlin, p 87–147, https://doi.org/10.5281/zenodo.1469555

Savary A, Liu J, Pierredon A, et al (2023) We thought the eyes of coreference were shut to Multiword Expressions and they mostly are. Journal of Language Modelling 11(1):147–187. URL https://jlm.ipipan.waw.pl/index.php/JLM/article/view/328

Svensson MH (2004) Critères de figement: L'Identification des Expressions Figées en Français Contemporain. PhD thesis, Moderna Språk

Taslimipoor S, Bahaadini S, Kochmar E (2020) MTLB-STRUCT@ parseme 2020: Capturing unseen Multiword Expressions using multi-task learning and pre-trained masked language models. arXiv preprint arXiv:201102541

Vincze V, Csirik J (2010) Hungarian Corpus of Light Verb Constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, China, pp 1110–1118, URL https://www.aclweb.org/anthology/C10-1125

Vincze V, Nagy T. I, Berend G (2011) Multiword Expressions and Named Entities in the Wiki50 Corpus. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. Association for Computational Linguistics, Hissar, Bulgaria, pp 289–295, URL https://www.aclweb.org/anthology/R11-1040

Zaghouani W, Diab M, Mansouri A, et al (2010) The revised Arabic Propbank. In: Proceedings of the fourth linguistic annotation workshop, pp 222–226