SAMER: A Semi-Automatically Created Lexical Resource for Arabic Verbal Multiword Expressions Tokens Paradigm and their Morphosyntactic Features

Mohamed Al-Badrashiny, Abdelati Hawwari, Mahmoud Ghoneim, and Mona Diab
Department of Computer Science, The George Washington University
{badrashiny,abhawwari,mghoneim,mtdiab}@gwu.edu

Abstract

Although *MWE* are relatively morphologically and syntactically fixed expressions, several types of flexibility can be observed in *MWE*, verbal *MWE* in particular. Identifying the degree of morphological and syntactic flexibility of *MWE* is very important for many Lexicographic and NLP tasks. Adding *MWE* variants/tokens to a dictionary resource requires characterizing the flexibility among other morphosyntactic features. Carrying out the task manually faces several challenges since it is a very laborious task time and effort wise, as well as it will suffer from coverage limitation. The problem is exacerbated in rich morphological languages where the average word in Arabic could have 12 possible inflection forms. Accordingly, in this paper we introduce a semi-automatic Arabic multiwords expressions resource (SAMER). We propose an automated method that identifies the morphological and syntactic flexibility of Arabic Verbal Multiword Expressions (*AVMWE*). All observed morphological variants and syntactic pattern alternations of an *AVMWE* are automatically acquired using large scale corpora. We look for three morphosyntactic aspects of *AVMWE* types investigating derivational and inflectional variations and syntactic templates, namely: 1) inflectional variation (inflectional paradigm) and calculating degree of flexibility; 2) derivational productivity; and 3) identifying and classifying the different syntactic types. We build a comprehensive list of *AVMWE*. Every token in the *AVMWE* list is lemmatized and tagged with POS information. We then search Arabic Gigaword and All ATBs for all possible flexible matches. For each *AVMWE* type we generate: a) a statistically ranked list of *MWE*-lexeme inflections and syntactic pattern alternations; b) An abstract syntactic template; and c) The most frequent form. Our technique is validated using a Golden *MWE* annotated list. The results shows that the quality of the generated resource is 80.04%.

1 Introduction

Multiword expressions (*MWE*) are complex lexemes that contain at least two words reflecting a single concept. They can be morphologically and syntactically fixed expressions but also we note that they can exhibit flexibility especially in verbal *MWE*. Such morphosyntactic flexibility increases difficulties in computational processing of *MWE* as they are harder to detect. Characterizing the internal structure of *MWE* is considered very important for many natural language processing tasks such as syntactic parsing and applications such as machine translation (Ghoneim and Diab, 2013; Carpuat and Diab, 2010). In lexicography, entries for *MWE* in a lexicon should provide a description of the syntactic behavior of the *MWE* constructions, such as syntactic peculiarities and morphosyntactic constraints (Calzolari et al., 2002). Automatically identifying the syntactic patterns and listing/detecting their possible variations would help in lexicographic representation of *MWE*, as the manual annotation of *MWE* variants suffer from many disadvantages such as time and effort consuming, subjectivity and limited coverage.

The problem is exacerbated for morphologically rich languages, where an average word could have up to 12 morphological analyses such as the case for the Arabic language which is highly inflectional.

Proceedings of the 12th Workshop on Asian Language Resources,
pages 113–122, Osaka, Japan, December 12 2016.

Several challenges are encountered in automatic identification and parsing of *MWE* in Arabic especially verbal ones, because of their highly morphosyntactic flexibility.

This paper focuses on the Arabic verbal *MWE*(*AVMWE*) in Modern Standard Arabic (MSA). We broadly consider a *MWE* as a verbal one if it contains at least one verb in its elements. We focus exclusively on the flexibility of the elements existing in the *AVMWE* and their syntactic alternatives. Lexical flexibility (word/word) is meant to be outside the scope of this paper (Ex. *rajaE bixuf∼ayo Hunayon*[1] Vs. *EAd bixuf∼ayo Hunayon* where both expressions mean **return empty handed**).

From a theoretical point of view, we identify four components, for each *AVMWE* as shown in Table 1. The verbal components are any verb within a *MWE*. Elements are the non-verbal components such as noun, adjective or particle. The syntactic variable is a slot that reflects the syntactic function in a *MWE* without being itself a part of the construction, and the gaps are some inserted modifiers that might occur between *MWE* elements (Hawwari et al., 2014).

	Verbal component	**Gap**	**Syntactic variable**	**Element-1**	**Element-2**	**Element-3**	**Syntactic variable**
BW	>aEoTaY	>amosi	(FulAnN)	AlDawo'	Al>axoDar	li-	(FulAnK)
En-Gloss	gave	yesterday	(somebody)	the-light	the-green	to	(something/somebody)
En-translation	(somebody) gave the green light to (somebody/something)						

Table 1: Example for the entities we are considering within a *MWE*

The main objective of our work is to automatically acquire all observed morphological variants and syntactic pattern alternations of a *MWE* using large scale corpora, using an empirical method to identify the morphological and syntactic flexibility of *AVMWE*.

2 Related Work

A considerable amount of literature has been published on the morphosyntactic characteristics of *MWE* . These studies focused on various morphological aspects, within different contexts on different languages. Gurrutxaga and Alegria (2012) and Baldwin et al. (2003) applied latent semantic analysis to build a model of multiword expression decomposability. This model measures the similarity between a multiword expression and its elements words, and considers the constructions with higher similarities are greater decomposability.

Diab and Bhutada (2009) present a supervised learning approach to classify the idiomaticity of the Verb-Noun Constructions (VNC) depending on context in running text.

Savary (2008) presents a comparative survey of eleven lexical representation approaches to the inflectional aspects in *MWE* in different languages, including English, French, Polish Serbian German, Turkish and Basque.

Al-Haj et al. (2013) applied to Modern Hebrew an architecture for lexical representation of MWEs. The goal was to integrate system that can morphologically process Hebrew multiword expressions of various types, in spite of the complexity of Hebrew morphology and orthography.

Zaninello and Nissim (2010) present three electronic lexical resources for Italian *MWE*. They created a series of example corpora and a database of *MWE* modeled around morphosyntactic patterns.

Nissim and Zaninello (2013) employed variation patterns to deal with morphological variation in order to create a lexicon and a repository of variation patterns for *MWE* in morphologically-rich Romance languages.

Al-Sabbagh et al. (2013) describe the construction of a lexicon of Arabic Modal Multiword Expressions and a repository for their variation patterns. They used an unsupervised approach to build a lexicon for Arabic Modal Multiword Expressions and a repository for their variation patterns. The lexicon contains 10,664 entries of MSA and Egyptian modal *MWE* and collocation, linked to the repository.

The closest work to ours is that of (Hawwari et al., 2012). They created a list of different types of Arabic *MWE* collected from various dictionaries which were manually annotated and grouped based on their syntactic type. The main goal was to tag a large scale corpus of Arabic text using a pattern-

[1]We use Buckwalter transliteration encoding for Arabic: http://www.qamus.org/transliteration.htm

matching algorithm and automatically annotated to enrich and syntactically classify the given *MWE* list. Their work didn't approach the derivational or lexical aspects.

To the best of our knowledge, to date, none of the previous addressed the systematic investigation of morphosyntactic features and derivational productivity of *AVMWE* and their syntactic properties.

3 Linguistic Background

This section gives a brief overview of the linguistic background of the verbal inflectional and derivational system in Modern Standard Arabic.

3.1 Arabic Verbal MWE (AVMWE)

The verbal *MWE* is a *MWE* that includes a verb or more within its word elements. *AVMWE* could be classified, according to their lexical nature, into three types:

- Verbal Idioms: We mean by verbal Idiom any idiomatic expression that has a verb within its components. An example of verbal idiom is as follows: *taraka (fulAnN) Al-jamala bi-maA Hamala*[2]. [**(someone) left every thing behind**]
- Light verb (support verb): a light verb construction is consisting of: a) a verb that is semantically light, and b) a noun or verbal-noun carries the core meaning of the construction. *>axa* (fulAnN) Al-v >ora* [**(someone take a revenge**]
- Verb Particle construction: An expression includes a verb and a particle that they have together a meaning. (this construction includes phrasal verbs): *ragiba (fulAnN) fi* [**wish for**]

A *MWE* is considered flexible when it has more than one accepted inflected or syntactic form. Flexibility can be applied to inflectional, derivational, syntactic and lexical aspects of a MWE. We roughly distinguish between flexibility and idiomaticity as follows: flexibility affects the morphosyntactic properties, and idiomaticity is more related to the compositionality and semantic content of an *MWE*.

Inflection is a morphological subfield that belongs to single words encoding its inflectional categories (number, gender, person, case, tense, voice, mood, aspect) using several affixes to represent the morphosyntactic variation. Inflectional flexibility of an *MWE* is a sum of the inflectional flexibility of its elements.

A *MWE* token instance includes every possible inflectional variation form of the *MWE* type that can occur in a corpus. On the other hand, a *MWE* type is the canonical (citation) form that is used to be the basic form representing all the possible tokens of a *MWE* lexeme. Lexicographers chose the simplest form to be a canonical form serving as a head word or citation form for a lexical entry. By an *MWE* lexeme we refer to all the possible inflectional forms that are observed for the *MWE* in a corpus.

3.2 Inflectional Categories

The Arabic verb has the following inflectional categories:

- Tense: perfective, imperfective, imperative
- Voice: active, passive
- Mood: indicative (*marofuwE*), subjunctive (*manoSuwb*), jussive (*majozuwm*)

However, verb subject inflects for person (first, second, third person), gender (masculine, feminine), number (singular, dual, plural) and syntactic case (nominative (*marofuwE*), accusative (*manoSuwb*), genitive (*majoruwr*)).

AVMWE vary in their inflectional flexibility degree. One group is fixed, for example *Had∼ivo wa-lA Haraj* (**speak freely**), second group has a degree of flexibility as *>aTolaq (fulAnN) sAqyohi li-AlryiH* (**ran away**), the verb *>aTolaq* is fully flexible for any affixes (*>aTlaqA, >aTlaquw, >aTlaqato*, etc).

[2]We use Buckwalter transliteration scheme to represent Arabic in Romanized script throughout the paper. http://www.qamus.org/transliteration.htm

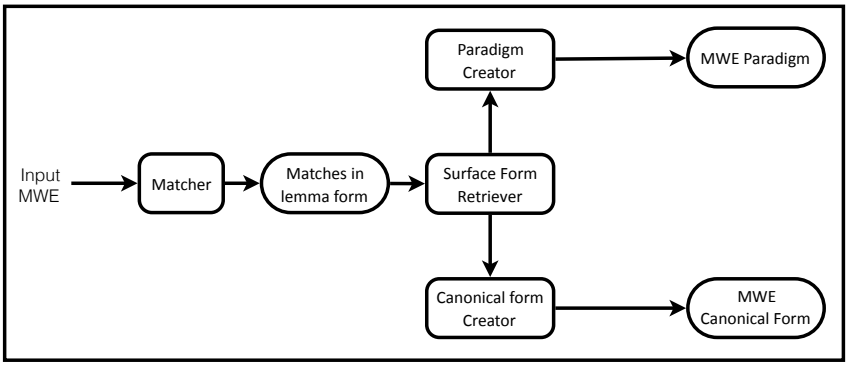

Figure 1: MWE paradigm and canonical form finder pipeline: MPD.

3.3 Derivational Productivity

Derivation is a very productive and regular word-formation mechanism in the Arabic language. Unlike inflection, derivation belongs to the lexicon and is completely syntax independent. As *AVMWE* vary in productivity, some allow verbal derivations, for example *fAza bi-qaSabi Alsaboqi* (**he came first/he is the winner**) allows the derived nominal *MWE fA}izN bi-qaSabi Alsaboqi*, where the verb *fAza* is derivationally related to the noun *fA}izN*. On the other hand, there are many *AVMWE* that are fixed derivationally as they do not exhibit derivational productivity for example *>aSobaH ha$iymaF ta*oruw-hu AlriyAHu* (**vanish**).

3.4 Syntactic flexibility

As it is a Verb Phrase (VP), *AVMWE* is governed by VP grammatical rules and operations such as word order, agreement, government. Syntactic flexibility for an *MWE* occurs in texts in different configurations. Some of *AVMWE* have some degree of syntactic flexibility, which appears in word order variability for a given *MWE* (VSO: *balag Alsayolu AlzubaY*(**it reached the limits**), SVO: *Alsayolu balag AlzubaY*). Although word order in Arabic is relatively free, the word-order flexibility in *AVMWE* occurs rarely, because the *AVMWE* phrases are more rigid than ordinary phrases syntactically. An example for the syntactic fixed *AVMWE* is *AixotalaTa AlHAbilu biAln∼Abili* (**it became a mess**).

4 Approach

We introduce an automatic approach for building a morphosyntactic lexicon for Arabic verbal *MWE* starting from a gold seed list. We use a manually created list of Arabic verbal *MWE* and try to find all possible matches with any morphological variations in a large dataset in a process of *MWE* Paradigm detection (MPD). After that we create the morphosyntactic feature vector of each match and calculate the level of flexibility of each *MWE*.

Figure 1 illustrates the different components of the MPD system. For each new *MWE* expression in seed list, the "Matcher" component replaces each word in the input *MWE* with its lemma to find all possible inflections for the *MWE* during the matching process. Since deverbals such as verbal nouns, past participle active, and past participle passive inherit the semantic and syntactic structures from their verbs they are derived from, the "Matcher" component adds the derivatives of each verb in the input *MWE* as possible matching candidates in addition to its lemma. Technically these are derivational variants. That way, we can find all possible forms of the input *MWE* during the matching process. For example, if the input is the *MWE* "*fAz biqaSabi Alsaboqi* meaning **he is the winner**", it will be matched with "*fAzuw biqaSabi Alsaboqi* meaning **the winning**" and "*fA}izN biqaSabi Alsaboqi* meaning **the winner**", reflecting inflectional variation with *fAz* being observed as *fAzuw* in the former, and derivational variation with *fAz* being observed as *fA}izN* in the latter.

The "Matcher" looks up the new form of the MWE (i.e. the lemma form with the different verb derivatives candidates) in large preprocessed datasets that are described in section 5.2 below while enabling any

possibility of gapping between the words (Ex. *fAz Alwalad biqaSabi Alsaboqi* **the boy is the winner**) or word ordering (Ex. *waqaEa fiy HaySa bayoSa*. Or *fiy HaySa bayoSa waqaEa* **got confused**).

The preprocessed datasets have a one-to-one mapping between the input surface form of each sentence and its corresponding lemma form. Thus, the "Surface Form Retriever" component uses that to find the original surface form of each sentence retrieved by the "Matcher" component. The "Paradigm Creator" component generates a unique list of all surface form sentences retrieved by the "Surface Form Retriever" component to create a list of all possible morphological variations of the input *MWE*. To make the list as generic as possible, we replace each word that is not part of the MWE with its POS tag. The "Canonical Form Creator" after that uses the full list of sentences created by the the "Surface Form Retriever" component and finds the most frequent matched form of the input *MWE*

For each word in the matched *MWE*, we create a morphosyntactic feature vector of nine elements that are being extracted from the POS tags of the matched *MWE*.The first element is the POS. Three elements (aspect, voice, mood) are only for verbs, while nominals have the following attributes (case, state), and (person, number, gender) apply to all words. In addition to that, we try to identify the candidate subject and object for each match as follows:

- Subject: The candidate subject is identified as the pronoun attached to the verb if it is explicitly mentioned in the pos of the verb, otherwise it is the first nominative nominal after the verb;
- Object: The candidate object is identified as the pronoun attached to the verb if it is explicitly mentioned in the pos of the verb. Otherwise it is the first accusative nominal after the verb.

5 Experimental Setup

5.1 Datasets

We use different types of datasets to evaluate our approach for creating the *MWE* token paradigm resource.

Corpora used for the resource creation:

- *ATB*: The Arabic Treebanks. The selected *ATBs* represent three different genres: Newswire;[3]Broadcast News;[4] and, Weblogs;[5]
- *Gigawords*: The Arabic Gigaword fourth edition;[6]
- *AVMWE*: Is a list of more about 4000 verbal *MWE* semi automatically extracted from two traditional Arabic Monolingual Dictionaries;
- *Verbs-to-Derivatives*: Is a list of 10k MSA verbs and their possible derivations. It is developed to help our system recognize the derivational relations between verbs and their nominal counterparts (Active participle, passive participle and Gerund) (Hawwari et al., 2013).

Evaluation Datasets:

- *DevDB*: 2000 randomly selected lines from the *ATB* and *Gigawords* used for system tuning;
- *TstDB*: 2000 randomly selected lines from the *ATB* and *Gigawords* used for system evaluation.

Both *DevDB* and *TstDB* are manually annotated. Each line is annotated with a presence/absence tag indicating whether an *MWE* from the *AVMWE* list or not. If a line is annotated as having an *MWE*, all of the elements of this *MWE* are annotated and the number of gaps between each two elements is identified. Table 2 shows the annotation distribution of both datasets

[3]ATB-P1-V4.1(LDC2010T13),ATB-P2-V3.1 (LDC2011T09) and ATB-P3-V3.2 (LDC2010T08)

[4]ATB-P5-V1.0 (LDC2009E72), ATB-P7-V1.0 (LDC2009E114), ATB-P10-V1.0 (LDC2010E22) and ATB-P12-V2.0 (LDC2011E17)

[5]ATB6-v1.0(LDC2009E108) and ATB11-v2.0(LDC2011E16)

[6]LDC2009T30

	Has-MWE	No-MWE
DevDB	42.85%	57.15%
TstDB	45.55%	54.45%

Table 2: *MWE* Annotation distribution across the evaluation datasets.

5.2 Data Preparation

To enable matching based on Lemma and POS, we processed the *ATB* and *Gigawords* into a a series of tuples with the following elements: "Token-Lemma-POS". For *ATB*, we extracted this format from the gold analysis in the integrated files. For *Gigawords*, we used *MADAMIRA* toolkit (Pasha et al., 2014) for tokenization, lemmatization and POS tagging. The selected tokenization scheme is ATB-tokenization and the POS tag-set is *ATB* full tag-set. The *AVMWE* was also processed using *MADAMIRA* to guarantee consistency in the matching process. *MADAMIRA* provides a list of possible analyses per word with the most probably one selected as the candidate analysis. Due to short context, the accuracy of the selected analysis by *MADAMIRA* wasn't high. Accordingly, we post-processed the list of possible analyses per word and selected the most probable analysis that matches the gold assigned coarse-grained POS.

6 Paradigm Detection Evaluation

We used the processed *AVMWE* list as the input gold *MWE* list to our paradigm detection system MPD. Also, *Verbs-to-Derivatives* is used to help the matching algorithm to match the derivatives of each verb in the input multi words expressions as well.

Table 3 shows the results of running the paradigm detector on *DevDB* with different schemes (i.e. different gapping sizes and with and without enabling word reordering). We report the results as the F-score of correctly tagging an *MWE* in *DevDB*, the F-score of correctly tagging the sentences that do not have *MWE*, and the weighted average F-score of both of them for all schemes. The results shows that the best weighted Average F-score is 80.61% when we allow a maximum gap size of 2 between the *MWE* constituent words and without enabling the word order to be varied.

By running the best setup on the *TstDB*, we found that the weighted average F-score is 80.04%

	with-words-reordering			**without-words-reordering**		
Max-Gap-Size	**MWE tagging**	**No-MWE tagging**	**Avg-Fscore**	**MWE tagging**	**No-MWE tagging**	**Avg-Fscore**
0	66.62%	81.75%	75.27%	65.80%	81.75%	74.92%
1	75.14%	82.34%	79.25%	73.81%	83.29%	79.23%
2	77.40%	80.39%	79.11%	77.09%	83.25%	**80.61%**
4	73.87%	70.30%	71.83%	76.20%	79.89%	78.31%
8	68.39%	52.53%	59.33%	73.42%	74.07%	73.79%
16	63.15%	26.02%	41.93%	69.94%	66.24%	67.83%
32	60.64%	5.93%	29.37%	68.04%	61.45%	64.27%
65	60.08%	0.70%	26.14%	67.82%	60.67%	63.73%
any	59.99%	0.00%	25.71%	67.82%	60.67%	63.73%

Table 3: F-score of correctly tagging the *MWE* in *DevDB* and the F-score of correctly tagging the sentences that do not have *MWE* with different experimental setups.

6.1 Error Analysis

Type	%
gap	31.23%
order	20.15%
pp-attachment	20.15%
polysemous	17.88%
MADAMIRA	6.55%
literal	3.53%
Eval-err	0.25%
Syn-function	0.25%

Table 4: Paradigm detector error analysis

Table 4 shows the error distribution of the paradigm detector on the *TstDB*. We can see that limiting the maximum gapping size to two and disabling word reordering while matching are the main sources of errors. Together they are responsible for 51.38% of the errors, which suggests that the gap size and word reordering should be more flexible. We should have some smarter way to decide the gapping size and words reordering status per *MWE* type; not by generalizing them on all types. For example "*ya>oxu* bi+ Eayon AlAiEotibAr* means **considers**" did not match with "***ta>oxu**** *mA TuriH fiy mu&otamar Al-manAmap* ***bi+ Eayon*** *Aljid∼iyap wa+* ***AlAiEotibAr***" because of the gapping size restriction. And "*ba*al jahodi +h* means **did his effort**" did not match with "***Aljuhuwd*** *Altiy* ***tabo*ulul +hA***" because the word reordering is disabled.

Another challenging problem responsible for 20.15% of the errors is the verb particle construction; where a certain verb when attached to a certain preposition, they act like an *MWE*. This issue is that while matching, it is hard to know if a certain preposition should be attached to the target verb or another one. This leads to false identification for the match if the decision of the attachment was not correct. Ex: "*yajib EalaY +h* means **he should be**" incorrectly matched with "***yajib*** *>n yaEoqid* ***EalaY >roDihi he has to held on his land***" because *EalaY* is considered attached to "*yajib*" while it is actually attached to *yaEoqid* as it assumed a gapping of two words, while it should have attached the particle to the low, second and closest verb *yaEoqid*.

Polysemy is also a hard problem. It is responsible for 17.88% of the errors. Errors due polysemy occur when words in the input *MWE* type have more than one meaning. But since the matching process only takes the lemma and POS into account and word senses are not part of the matching, the paradigm detector could tag some cases as valid matches. Ex:"*Hayovu kAn* meaning **wherever**" is incorrectly matched with "***Hayovu kAn*** *AlAibonu yaloEab* meaning **because the sone was playing**". The issue came from the word *Hayovu* that means **where** or **because**.

The morphological analyzer and POS tagger (*MADAMIRA*) is the source of 6.55% of the errors. When *MADAMIRA* incorrectly analyzes some words, some wrong matches occur. Ex: "**ahabat riyHu +hu* means **has been forgotten**" did not match with "****ahabot riyHi +hu***" because *MADAMIRA* analyzed the word "**ahabat* means **gone**" as "**ahabot* means **I went**"

3.53% of the errors are due to the *MWE* being idiomatic in some contexts and literal in others. Ex. "*tajAwaz Huduwd +hu*" meaning "**Exceeded his limits**" incorrectly matched "***tatajAwaz AlHuduwd*** *AljugorAfiy∼ap*" meaning "**Transcended the geographic boundaries**"

The remaining 0.5% errors are due to some minor issues: 0.25% errors are due to manual annotation errors, while the other 0.25% errors are due to fact that the matched morphological variant from the input *MWE* has a different syntactic function than the input *MWE*. Ex. "*HAwal EabavAF*" meaning "**Tried in vain**" is incorrectly matched with "***yHAwl AlEbv***" meaning "**Attempted to tamper with**". This is because the word "*EabavAF*" which is an adverb is a derivation of the noun "*AlEbv*" which plays the role of an object in this verb noun construction.

7 SAMER

To build the proposed Arabic *MWE* resource, we ran the paradigm detector on the *ATB* and *Gigawords* using the best configuration we found. The system found 732335 matches for 1884 *MWE* out of the 4000

MWE in the input *AVMWE* list.

The automatically created resource is reflected in the following five tables:

- All matches table: Contains the 732335 matches that are automatically detected by the paradigm detector and pointers to their original locations in the *ATB* and *Gigawords*;
- Flexibility table: This table has the 1884 rows representing the types of *MWE* that the paradigm detector found matches for. The columns represent the words of the *MWE* where the value of each cell shows the number of different forms that this element matched with. For example if a certain cell has the number "5", this means that its corresponding word matched with five different unique morphological variants;
- MWE-Lexeme table: This table shows the different morphological forms of each word in each *MWE* and their probabilities that are identified by the paradigm detector;
- Sorted-Grouped-tokens table: Shows the probability of all matches of each *MWE* in a descending order. So, if there is a *MWE* that has 10 matches, we calculate the unique form for each of them and find the probability of each unique value. The number of grouped types of all matches is 38408;
- MWE-Types table: this table has1884 rows; one row for each *MWE* type. The columns show number of matches, the most frequent token with its probability, and the union of the morphosyntactic features of each word across all tokens of each *MWE* type. Example: if the union of the gender of the second word across all matches of *MWE* number i is $\{$M,F$\}$; this means that the second word of the *MWE* number i has a flexibility to change the gender between masculine or feminine.

7.1 Statistical Analyses

The number of the *MWE* types in our automatically created resource is 1884. They consist of 1901 unique verbal words and 3104 unique non-verbal words. Each type of the 1884 *MWE* has an average fan out of 20 different forms due to the morphological or inflectional changes the *MWE* words.

The results show that 15.5% of the *MWE* types do not allow any gaps between the constituent words (No-Gaps), while 52.1% of the *MWE* types allow gapping between all the constituent words (Full-Gaps) and the remaining 32.4% of the types allow gapping only between some of the constituent words (Part-Gaps).

Examples:

- No-Gaps: "*dub∼ira bi+ layolK* meaning **conspired**" matched with "*dub∼ira bi+ layolK*"
- Full-Gaps: "*ka$af AlqinAE Ean* meaning **unveiled** " matched using one gap between the first two words with "*ka$af b +h AlqnAE En* meaning **unveiled using it**" and using one gap between the second two words with "*tk$f AlqnAE AlzA}f En* meaning **unveiled the fake thing**"
- Part-Gaps: "*ka$∼ar Ean >anoyAbi +h* meaning **express anger**" matched using one gap between the first two words with "*tuka$∼ir turokiyA Ean >anoyAbi +hA* meaning **Turkey expressed its anger**"

We found that 15.7% of the *MWE* types are fixed. They do not have any morphological or inflectional variations in all matched instances (Ex: *lA yaxoTuro bi+ bAlK* meaning **it will never come to your mind**). But the other 84.4% have a higher degree of flexibility that they can match with instances with different morphological or inflectional variations (Ex: *HAla duwna* that means "**prevented**" has a match with *tHwl duwna*). 4.7% of the matched verbal *MWE* types have matches with the derivatives of the verbal part (Ex: *kAl bi+ mikoyAlayon* meaning "**injustice**" is matched with *Alkyl bi+ mikoyAlayon*). Furthermore, the results show that non-verbal components of the *MWE* type have more tendency to stay fixed than the verbal parts. Since 51.7% of the non-verbal components stay fixed in all matched instances while only 17.7% of the verbs stay fixed.

Tables 5 and 6 show the morphosyntactic feature flexibility distribution for the non-verbal components and the verbal ones respectively across all *MWE* matches. The tables show that the mood is the most rigid feature (76.4% of the *MWE* types have fixed mood) while gender is the most flexible feature (87.08% of the *MWE* types have different values of the gender within the matched cases).

Feature	Fixed	Flexible
gender	87.08%	12.92%
number	85.18%	14.82%
case	56.28%	43.72%
state	63.66%	36.34%

Table 5: Morphosyntactic feature flexibility of the non-verbal components of all *MWE* types

Feature	Fixed	Flexible
aspect	27.7%	72.3%
voice	82.9%	17.1%
mood	23.6%	76.4%

Table 6: Morphosyntactic feature flexibility of the verbal components of all *MWE* types

8 Conclusion

We introduced an automatically built *MWE* resource that covers all the morphological variations of a list of *AVMWE* in the basic form. Each morphological variant is accompanied with all of its instances in the *ATB* and Arabic *Gigawords*. Furthermore, for each word in the *MWE*, we added a morphosyntactic feature vector of nine elements {pos, aspect, voice, mood, person, gender, number, case, state)}. We validated our approach constructing an automatic *MWE* paradigm detector in running text. Our system yielded an weighted average f-score of 80.61% on a dev set, and 80.04% on an unseen test data. The error analysis shows that there is no generalized maximum gapping size, and enabling or disabling word reordering decisions should not be generalized on all *MWE* in the input list. Instead, more sophisticated techniques are required to find the best decisions for each case.

References

Hassan Al-Haj, Alon Itai, and Shuly Wintner. 2013. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, page ect036.

Rania Al-Sabbagh, Jana Diesner, and Roxana Girju. 2013. Using the semantic-syntactic interface for reliable arabic modality annotation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 410–418, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicoletta Calzolari, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *LREC*.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 242–245, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 17–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mahmoud Ghoneim and Mona T Diab. 2013. Multiword expressions in the context of statistical machine translation. In *IJCNLP*, pages 1181–1187.

Antton Gurrutxaga and Iaki Alegria. 2012. Measuring the compositionality of nv expressions in basque by means of distributional similarity techniques. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet U?ur Do?an, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Abdelati Hawwari, Kfir Bar, and Mona Diab. 2012. Building an arabic multiword expressions repository. *Proc. of the 50th ACL*, pages 24–29.

Abdelati Hawwari, Wajdi Zaghouani, Tim O'Gorman, Ahmed Badran, and Mona Diab. 2013. Building a lexical semantic resource for arabic morphological patterns. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6, Feb.

Abdelati Hawwari, Mohammed Attia, and Mona Diab. 2014. A framework for the classification and annotation of multiword expressions in dialectal arabic. *ANLP 2014*, page 48.

Malvina Nissim and Andrea Zaninello. 2013. Modeling the internal variability of multiword expressions through a pattern-based method. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2):7.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of LREC*, Reykjavik, Iceland.

Agata Savary. 2008. Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2):1–53.

Andrea Zaninello and Malvina Nissim. 2010. Creation of lexical resources for a characterisation of multiword expressions in italian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).