Unsupervised Construction of a Lexicon and a Repository of Variation Patterns for Arabic Modal Multiword Expressions

Rania Al-Sabbagh†, Roxana Girju†, Jana Diesner‡
†Department of Linguistics and Beckman Institute
‡School of Library and Information Science
University of Illinois at Urbana-Champaign, USA
{alsabba1, girju, jdiesner}@illinois.edu

Abstract
We present an unsupervised approach to build a lexicon of Arabic Modal Multiword Expressions (AM-MWEs) and a repository of their variation patterns. These novel resources are likely to boost the automatic identification and extraction of AM-MWEs.

1 Introduction
Arabic Modal Multiword Expressions (AM-MWEs) are complex constructions that convey (un)certainty, evidentiality, obligation, permission, commitment, ability and volition.

AM-MWEs range from completely fixed, idiomatic and sometimes semantically-opaque expressions, to morphologically, syntactically and/or lexical productive constructions. As a result, the identification and extraction of AM-MWEs have to rely on both a lexicon and a repository of their variation patterns. To-date and to the best of our knowledge, neither resource is available. Furthermore, AM-MWEs are quite understudied despite the extensive research on general-purpose Arabic MWEs.

To build both the lexicon and the repository, we design a four-stage unsupervised method. Stage 1, we use Log-Likelihood Ratio and a root-based procedure to extract candidate AM-MWEs from large Arabic corpora. Stage 2, we use token level features with k-means clustering to construct two clusters. Stage 3, from the clustering output we extract patterns that describe the morphological, syntactic and semantic variations of AM-MWEs, and store them in the pattern repository. Stage 4, we use the most frequent variation patterns to bootstrap low-frequency and new AM-MWEs. The final lexicon and repository are manually inspected. Both resources are made publicly available.

The contributions of this paper are: (1) we address the lack of lexica and annotated resources for Arabic linguistic modality; and hence, we support NLP applications and domains that use modality to identify (un)certainty (Diab et al. 2009), detect power relations (Prabhakaran and Rambow 2013), retrieve politeness markers (Danescu-Niculescu-Mizil et al. 2013), extract and reconstruct storylines (Pareti et al. 2013) and classify request-based emails (Lampert et al. 2010); (2) we provide both a lexicon and a repository of variation patterns to help increase recall while keeping precision high for the automatic identification and extraction of productive AM-MWEs; and (3) we explore the morphological, syntactic and lexical properties of the understudied AM-MWEs.

For the rest of this paper, Section 2 defines AM-MWEs. Section 3 outlines related work. Sections 4 describes our unsupervised method. Section 5 describes manual verification and the final resulting resources.

2 What are AM-MWEs?
AM-MWEs are complex constructions that convey (un)certainty, evidentiality, obligation, permission, commitment, ability and volition. Based on their productivity, we define five types of AM-MWEs:

Type 1 includes idiomatic expressions like

\[\text{HtmA w/Abd} \text{ (must),} \quad \text{I/EI w/EsY} \quad \text{(maybe) and} \quad \text{yrgb fy} \quad \text{(he wants to)} \quad \text{and} \quad \text{wAvq mn} \quad \text{(sure about)}. \]

They inflect...
there is). In

Table 1: Example AM-MWEs and their unigram synonyms

<table>
<thead>
<tr>
<th>AM-MWEs</th>
<th>Unigram Synonym(s)</th>
<th>English Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eqdr AlEzn  ElY</td>
<td>-</td>
<td>I intended to</td>
</tr>
<tr>
<td>yrkm An</td>
<td>-</td>
<td>I can/I have the ability to</td>
</tr>
<tr>
<td>hnyk ADzqAd bAn</td>
<td>-</td>
<td>I think</td>
</tr>
<tr>
<td>hnya AItmAIt bAn</td>
<td>-</td>
<td>possibly/there is a possibility that</td>
</tr>
</tbody>
</table>

for gender, number, person, and possibly for tense, mood and aspect. Neither the head word nor the preposition is replaceable by a synonym. In the literature of MWEs, Type 2 is referred to as phrasal verbs. In the literature of modality, it is referred to as quasi-modalities (i.e. modalities that subcategorize for prepositions).

**Type 3** comprises lexically productive expressions whose meanings rely on the head noun, adjective or verb. If the head word is replaced by another of the same grammatical category but a different meaning, the meaning of the entire expression changes. Hence, if we replace the head adjective AlDrwry (necessary) in

\[
\text{mn AlDrwry An (it is necessary to) with Almkn (possible), the meaning changes from obligation to uncertainty.}
\]

**Type 4** comprises syntactically productive expressions. It is similar to Type 3 except that the head words are modifiable and their arguments, especially indirect objects, can be included within the boundaries of the MWE. Thus, the same expression from Type 3 can be modified as in

\[
\text{mn AlDrwry jDA An (it is necessary to). Furthermore, we can have an inserted indirect object as in}
\]

\[
\text{mn AlDrwry lImSryn An (it is necessary for Egyptians to).}
\]

**Type 5** includes morphologically, lexically and syntactically productive expressions like

\[
\text{ldy yqyn An (I have faith that). Morphologically, the object pronoun in ldy (I have) inflects for person, gender and number. Syntactically, the head noun can be modified by adjectives as in}
\]

\[
\text{ldy yqyn rAxx An (I have a strong faith that). Lexically, the meaning of the expression relies on the head noun yqyn (faith) which is replaceable for other modality-based nouns such as nyp An (I have an intention to).}
\]

Despite the semantic transparency and the morpho-syntactic and lexical productivity of the expressions in Types 3-5, we have three reasons to consider them as AM-MWEs:

First, although the head words in those expressions are transparent and productive, the other components, including prepositions, relative adverbials and verbs, are fixed and conventionalized. In

\[
\text{mn AlDrwry An (literally: from the necessary to; gloss: it is necessary to), the preposition mn (from) cannot be replaced by any other preposition. In}
\]

\[
\text{hnAk AHtmAl bAn (there is a possibility that), the relative adverbial hnAk (there is) cannot be replaced by another relative adverbial such as hna (there is). In}
\]

\[
\text{hDwny AIamf fy An (hope derives me to), the head is the noun AIamf (the hope). Therefore, the lexical verb hDwny (drives me) cannot be replaced by other synonymous verbs such as yqdn (leads me) or ydfEny (pushes/drives me).}
\]

Second, each of those expressions has a strictly fixed word order. Even for expressions that allow the insertion of modifiers and verb/noun arguments, the inserted elements hold fixed places within the boundaries of the expression. Complex constructions that adhere to strict constraints on word order but undergo lexical variation are classified by Sag et al. (2002) as semi-fixed MWEs.

Finally, each expression of those types is lexically perceived as a one linguistic unit that can be replaced in many contexts by a unigram synonym as illustrated in Table 1. According to Stubbs (2007) and Escartín et al. (2013), the perception of complex constructions as single linguistic units is characteristic of MWEs.

### 3 Related Work

There is a plethora of research on general-purpose Arabic MWEs. Yet, no prior work has focused on AM-MWEs. Hawwari et al. (2012) describe the manual construction of a repository for Arabic MWEs that classifies them based on their morpho-syntactic structures.
Attia et al. (2010) describe the construction of a lexicon of Arabic MWEs based on (1) correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages, (2) English MWEs extracted from Princeton WordNet 3.0 and automatically translated into Arabic, and (3) lexical association measures.

Bounhas and Slimani (2009) use syntactic patterns and Log-Likelihood Ratio to extract environmental Arabic MWEs. They achieve precision rates of 0.93, 0.66 and 0.67 for bigrams, trigrams and quadrigrams, respectively.

Al-Sabbagh et al. (2013) manually build a lexicon of Arabic modals with a small portion of MWEs and quasi-modals. In this paper, quasi-modals are bigram AM-MWEs. Hence, their lexicon has 1,053 AM-MWEs.

Nissim and Zaninello (2013) build a lexicon and a repository of variation patterns for MWEs in the morphologically-rich Romance languages. Similar to our research, their motivation to represent the productivity of Romance MWEs through variation patterns is to boost their automatic identification and extraction. Another similarity is that we define variation patterns as part-of-speech sequences. The difference between their research and ours is that our variation patterns have a wider scope because we cover both the morpho-syntactic and lexical variations of AM-MWEs, whereas their variation patterns deal with morphological variation only.

4 The Unsupervised Method

4.1 Extracting AM-MWEs

4.1.1 Extraction Resources

Table 2 shows the token and type counts as well as the descriptions of the corpora used for extraction. For corpus preprocessing, (1) html mark-up and diacritics are removed. (2) Meta-

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Token #</th>
<th>Types #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajdir</td>
<td>113774517</td>
<td>2217557</td>
<td>a monolingual newswire corpus of Modern Standard Arabic</td>
</tr>
<tr>
<td>LDC ISI</td>
<td>28880558</td>
<td>522443</td>
<td>an LDC parallel Arabic-English corpus (Munteanu &amp; Marcu 2007)</td>
</tr>
<tr>
<td>YADAC</td>
<td>6328248</td>
<td>457361</td>
<td>a dialectal Arabic corpus of Weblogs and tweets (Al-Sabbagh &amp; Girju 2012)</td>
</tr>
<tr>
<td>Tashkeel</td>
<td>596513</td>
<td>359050</td>
<td>a vowelized corpus of Classical and Modern Standard Arabic books</td>
</tr>
<tr>
<td>Total</td>
<td>41472307</td>
<td>3566311</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Statistics for the extraction corpora

We restrict the size of AM-MWEs in this paper to quadrigrams. Counted grams include function and content words but not affixes. Working on longer AM-MWEs is left for future research.

The extraction of candidate AM-MWEs is conducted in three steps:

**Step 1**: we use root-based information to identify the words that can be possible derivations of modality roots. For modality roots, we use the Arabic Modality Lexicon from Al-Sabbagh et al. (2013).

In order to identify possible derivations of modality roots, we use RegExps. For instance, we use the RegExp (\(w^*m(w^*)k(w^*)\) to identify words such as *Alumkna* (the possible), *Amkna* (I manage) and *bAmkAny* (I can) which convey modality.

This RegExp-based procedure can result in noise. For instance, the aforementioned RegExp also returns the word *الأمريكَان* *AlAmrykAn* (Americans) which happens to have the same three letters of the root in the same order although it is not one of its derivations. Yet, the procedure still filters out many irrelevant words that have nothing to do with the modality roots.

**Step 2**: for the resulting words from Step 1, we extract bigrams, trigrams and quadrigrams given the frequency thresholds of 20, 15 and 10, respectively.

2Ajjir: http://aracorpus.e3rab.com/  
Tashkeel: http://sourceforge.net/projects/tashkeela/
In previous literature on MWEs with corpora of 6-8M words, thresholds were set to 5, 8 and 10 for MWEs of different sizes. Given the large size of our corpus, we decide to use higher thresholds.

**Step 3:** for the extracted ngrams we use the Log-Likelihood Ratio (LLR) to measure the significance of association between the ngram words. LLR measures the deviation between the observed data and what would be expected if the words within the ngram were independent. Its results are easily interpretable: the higher the score, the less evidence there is in favor of concluding that the words are independent.

LLR is computed as in Eq. 1 where $O_{ij}$ and $E_{ij}$ are the observed and expected frequencies, respectively. LLR is not, however, the only measure used in the literature of MWEs. Experimenting with more association measures is left for future work.

**Eq. 1:**

$$\text{LLR} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Table 3 shows the unique type counts of the extracted ngrams. The extracted ngrams include both modal and non-modal MWEs. For instance, both $mn \text{ Almmkn lnA An}$ (it is possible for us to) and $fy \text{ Aqr}b \text{ wqt mmkn}$ (as soon as possible) are extracted as valid quadrigrams. Both have the word $mmkn$ (possible) derived from the root $m$-$k$-$n$. Both are frequent enough to meet the frequency threshold. The words within each quadrigram are found to be significantly associated according to LLR. Nevertheless, $mn \text{ Almmkn lnA An}$ is an AM-MWE according to our definition in Section 2, but $fy \text{ Aqr}b \text{ wqt mmkn}$ is not. This is because the former conveys the modality sense of possibility; whereas the latter does not. Therefore, we need the second clustering stage in our unsupervised method to distinguish modal from non-modal MWEs.

<table>
<thead>
<tr>
<th>Ngram size</th>
<th>Unique Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigrams</td>
<td>86645</td>
</tr>
<tr>
<td>Trigrams</td>
<td>43397</td>
</tr>
<tr>
<td>Quadrigrams</td>
<td>25634</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>96031</strong></td>
</tr>
</tbody>
</table>

Table 3: Statistics for the extracted MWEs

3 We use Banerjee and Pedersen's (2003) Perl implementation of ngram association measures.

### 4.2 Clustering AM-MWEs

Clustering is the second stage of our unsupervised method to build the lexicon of the AM-MWEs and the repository of their variation patterns. This stage takes as input the extracted ngrams from the first extraction stage; and aims to distinguish between the ngrams that convey modality senses and the ngrams that do not.

#### 4.2.1 Clustering Set-up

The clustering feature set includes token level morphological, syntactic, lexical and positional features. It also has a mixture of nominal and continuous-valued features as we explain in the subsequent sections.

#### 4.2.1.1 Morphological Features

Roots used to guide the extraction of candidate AM-MWEs in Section 4.1.2 are used as clustering morphological features. The reason is that some roots have more modal derivations than others. For instance, the derivations of the root $D-r-r$ include $Drwry$ (necessary), $bAlDrwrp$ (necessarily), and $DTTr$ (he has to); all of which convey the modality sense of obligation. Consequently, to inform the clustering algorithm that a given ngram was extracted based on the root $D-r-r$ indicates that it is more likely to be an AM-MWE.

#### 4.2.1.2 Syntactic Features

In theoretical linguistics, linguists claim that Arabic modality triggers (i.e. words and phrases that convey modality senses) subcategorize for clauses, verb phrases, to-infinitives and deverbal nouns. For details, we refer the reader to Mitchell and Al-Hassan (1994), Brustad (2000), Badawi et al. (2004) and Moshref (2012).

These subcategorization frames can be partially captured at the token level. For example, clauses can be marked by complementizers, subject and demonstrative pronouns and verbs. To-infinitives in Arabic are typically marked by $An$ (to). Even deverbal nouns can be detected with some POS tagsets such as Buckwalter's (2002) that labels them as NOUN.VN.

Based on this, we use the POS information around the extracted ngrams as contextual syntactic features for clustering. We limit the
window size of the contextual syntactic features to ±1 words.

Furthermore, as we mentioned in Section 2, we define AM-MWEs as expressions with fixed word order. That is, the sequence of the POS tags that represent the internal structure of the extracted ngrams can be used as syntactic features to distinguish modal from non-modal MWEs.

4.2.1.3 Lexical Features
As we mentioned in Section 2, except for the head words of the AM-MWEs, other components are usually fixed and conventionalized. Therefore, the actual lexical words of the extracted ngrams can be distinguishing features for AM-MWEs.

4.2.1.4 Positional Features
AM-MWEs, especially trigrams and quadrigrams that scope over entire clauses, are expected to come in sentence-initial positions. Thus we use @beg (i.e. at beginning) to mark whether the extracted ngrams occur at sentence-initial positions.

4.2.1.5 Continuous Features
Except for nominal morphological and lexical features, other features are continuous. They are not extracted per ngram instance, but are defined as weighted features across all the instances of a target ngram.

Thus, @beg for ngram, is the probability of ngram, to occur in a sentence-initial position. It is computed as the frequency of ngram occurring at a sentence-initial position normalized by the total number n of ngram, in the corpus.

Similarly, POS features are continuous. For instance, the probability that ngram is followed by a deverbal noun is the frequency of its POS, normalized as a deverbal noun tagged as a deverbal noun normalized by the total number n of ngram, in the corpus.

4.2.2 Clustering Resources
As we mentioned earlier, the extracted ngrams from the extraction stage are the input for this clustering stage. The root features are the same roots used for extraction. The POS features are extracted based on the output of MADAMIRA (Pasha et al. 2014) that is used to preprocess the corpus - Section 4.1.1. The positional features are determined based on the availability of punctuation markers for sentence boundaries.

We implement k-means clustering with k set to two and the distance metric set to the Euclidean distance. The intuition for using k-means clustering is that we want to identify AM-MWEs against all other types of MWEs based on their morpho-syntactic, lexical and positional features. Thus the results of k-means clustering with k set to two will be easily interpretable. Other clustering algorithms might be considered for future work.

4.2.3 Clustering Evaluation and Results
4.2.3.1 Evaluation Methodology
We use precision, recall and F1-score as evaluation metrics, with three gold sets: BiSet, TriSet and QuadSet, for bigrams, trigrams and quadrigrams, respectively. Each gold set has 1000 positive data points (i.e. AM-MWEs).

The gold sets are first compiled from multiple resources, including Mitchell and Al-Hassan (1994), Brustad (2000), Badawi et al. (2004) and Moshref (2012). Second, each compiled gold set is further evaluated by two expert annotators. They are instructed to decide whether a given ngram is an AM-MWE or not according to the following definitions of AM-MWEs:

- They convey modality senses - Section 1
- They have unigram synonyms
- They have fixed word orders
- Their function words are fixed

Inter-annotator kappa κ scores for the BiSet, TriSet and QuadSet are 0.93, 0.95 and 0.96, respectively. Most disagreement is attributed to the annotators' failure to find unigram synonyms.

The positive BiSet includes (1) phrasal verbs such as ymkmn nn (he manages to), yEjz En (he fails to) and yHlm be (he longs for), (2) prepositional phrases such as mn Almmkn (it is possible that) and fy AlHqyq (actually), (3) nominal phrases such as Amln hwe (my hope is to) and (4) AM-MWEs subcategorizing for complementizers such as SrH yErf An (he declares that) and yErf An (he knows that).

---

4 We use the k-means clustering implementation from Orange toolkit http://orange.biolab.si/
The positive TriSet includes verb phrases like 

(he fails to) and prepositional phrases like (it is impossible to) and Endy AymAn bAn

(I have faith that).

The positive QuadSet includes verb phrases such as 

(hope drives me to) and prepositional phrases such as 

and the latter misses the first preposition

(literally: in; gloss: have).

With these gold sets, we first decide on the best cluster per ngram size. We use an all-or-nothing approach; that is, for the two clusters created for bigrams, we select the cluster with the highest exact matches with the BiSet to be the best bigram cluster. We do the same thing for the trigram and quadgram clusters. With information about the best cluster per ngram size, our actual evaluation starts.

To evaluate clustered bigram AM-MWEs, we consider the output of best bigram, trigram and quadgram clusters to allow for evaluating bigrams with gaps. We also tolerate morphological differences in terms of different conjugations for person, gender, number, tense, mood and aspect.

For example, true positives for the bigram AM-MWE 

are only considered if the exact match and the morphological alternations of 

are included. In other words, if the output of the bigram clustering has Atmkn mn or ntmkn mn but the BiSet has only ytmkn mn, we consider this as a true positive.

The bigram ytmkn mn can have a (pro)noun subject after the verb ytmkn: ytmkn ((pro)noun gap) mn. Thus, we consider the output of the trigram best cluster. If we find ytmkn ((pro)noun gap) mn, we consider this as a true positive for the bigram ytmkn mn as long as the two boundaries of the bigrams are represented. We could not go any further with more than two gaps because we did not cluster beyond quadigrams.

False positives for the bigram ytmkn mn would be the bigrams ytmkn Alr}ys (the president manages) and the president from Alr}ys mn (the president to) in the bigram cluster where one of the bigram's components - either the verb or the preposition - is missing.

False negatives of bigrams would be those bigrams that could not be found in any of the best clusters whether with or without gaps.

Similar to evaluating bigrams, we consider the output of the trigram and quadgram best clusters to evaluate trigram AM-MWEs. We also tolerate morphological productivity.

For instance, the trigram 

endy aymAn bAn (I have faith that), that is conjugated for the first person plural is a true positive for the gold set trigram 

Endy AymAn bAn (I have faith that), is conjugated for the first person singular.

The same trigram Endy AymAn bAn can have two types of gaps. The first can be a noun-based indirect object after the preposition End. Thus, we can have

Endy AymAn bAn mTlq bAn (I have a strong faith that).

Consequently, in the output of the quadgram best cluster, if we find matches to Endy AymAn (adjective gap) bAn in any conjugations of Endy, or if we find any matches for End (noun gap) AymAn bAn, we consider them as true positives for the trigram Endy AymAn bAn.

If the pronoun in End is replaced by a noun and the adjective gap is filled, we will have a pentagram like 

End AlnAs AymAn mTlq bAn (people have a strong faith that). Since we do not extract pentagrams, we consider chunks such as 

End AlnAs AymAn (people have faith) and AymAn mTlq bAn (strong faith that) as false positive trigrams. This is because the former misses the complementizer bAn (in that), and the latter misses the first preposition End (literally: in; gloss: have).
Since we do not cluster pentagrams, we could not tolerate gaps in the output of the quadrigrams. We, however, tolerate morphological variation. As a result, \textit{yHdwnA AlAmI fy An} (hope drives us to) is considered as a true positive for \textit{yHdwny AlAmI fy An} (hope derives me to).

It is important to note that we do not consider the next best cluster of the larger AM-MWEs unless we do not find any true positives in the AM-MWE's original cluster. For example, we do not search for bigrams' true positives in the trigram and quadrigram clusters, unless there are not any exact matches of the gold-set bigrams in the bigrams' best cluster itself. The same thing applies when evaluating trigram AM-MWEs.

### 4.2.3.2 Clustering Results and Error Analysis

Table 4 shows the evaluation results for bigrams, trigrams and quadrigrams. We attribute the good results to our evaluation methodology in the first place because it allows counting true positives across clusters of different ngram sizes to account for gaps and tolerates morphological variations. Our methodology captures the morphological productivity of AM-MWEs which is expected given that Arabic is morphologically-rich. It also accounts for the syntactic productivity in terms of insertion.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigrams</td>
<td>0.663</td>
<td>0.776</td>
<td>0.715</td>
</tr>
<tr>
<td>Trigrams</td>
<td>0.811</td>
<td>0.756</td>
<td>0.783</td>
</tr>
<tr>
<td>Quadrigrams</td>
<td>0.857</td>
<td>0.717</td>
<td>0.780</td>
</tr>
</tbody>
</table>

Table 4: Clustering evaluation results

Long dependencies are a source of errors at the recall level. Clustering could not capture such instances as \textit{Alr}ys \textit{AlmSry Hsny mbArk b} (the Egyptian president Hosni Mubarak declared to) because they go beyond our quadrigram limit.

Another type of recall errors results from AM-MWEs that do not meet the extraction frequency threshold despite the large size of our corpus. Our positive gold sets are sampled from theoretical linguistics studies in which the included illustrative examples are not necessarily frequent. For example, we could not find instances for the volitive \textit{ytwq Aly} (he longs for).

Precision errors result from the fact that our RegExp-based procedure to guide the first extraction stage is noisy. For instance, the RegExp \((w^*)((w^*)w^*g(w^*))\) that was supposed to extract the volitive \textit{ytwq} (he longs) did not return any instances for the intended modal but rather instances for \textit{ytwqf} (he stops) which interestingly subcategorizes for a preposition and a complementizer as in \textit{ytwqf En An} (literally: stops from to). This subcategorization frame is the same for modals such as \textit{yEjz En An} (literally: unable from to). Consequently, \textit{ytwqf En An} (he stops from to) has been clustered as a trigram AM-MWE although it does not convey any modality senses. This highlights another reason for precision errors. The subcategorization frames and hence the syntactic features used for clustering are not always distinctive for AM-MWEs.

The \textit{@beg} feature was the least informative among all features. In the case of bigrams, they are mostly lexical verbs that do not occur in sentence initial positions. Meanwhile, punctuation inconsistencies do not enable us to reliably mark \textit{@beg} for many ngrams.

### 4.3 Identifying Variation Patterns

Our target is to build a lexicon and a repository of the variation patterns for AM-MWEs to boost their automatic identification and extraction, given their morpho-syntactic and lexical productivity.

In order to identify variation patterns, we use as input the best clusters from the previous clustering stage and follow these steps:

- We keep all function words as is with their lexical and POS representations
- We collapse all morphological tags for gender, number, person, tense, mood, aspect and case
- We add a \textit{HEAD} tag to the head words (i.e. words whose roots were used for extraction)
- We add a \textit{GAP} tag for adverbs, pronouns and other gap fillers to explicitly mark gap locations

An example pattern for the root \textit{- T-m-H} (wish) is \(((\text{HEAD}/IV^*) + (AL/Y/PREP) + (AN/SUB/CONJ))\) which reads as follows: a
trigram AM-MWE whose head is a verb in any conjugation followed by the preposition َيَلَوْنَ (to) and the subordinate conjunction َذَا (that; to). Another pattern that results from the aforementioned steps for the same root of تَمُّنُـهُ is (HEAD/*IV*) + (ADV/GAP) + (َيَلَوْنَ/PREP) + (َذَا/SUB_CONJ)). It means that an adverb can be inserted in-between the HEAD and the preposition َيَلَوْنَ (to).

4.4 Bootstrapping AM-MWEs

We use the patterns identified in the previous stage in two ways: first, to extract low-frequency AM-MWEs whose HEADs have the same roots as the pattern’s HEAD; and second, to extract AM-MWEs that have the same lexical, POS patterns but are not necessarily derived from the modality roots we used in extraction.

For example, from the previous section we used (HEAD/*IV*) + (َيَلَوْنَ/PREP) + (َذَا/SUB_CONJ)) to extract the third person feminine plural conjugation of the root تَمُّنُـهُ in the trigram يَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (they wish for) that occurred only once in the corpus. We used the same pattern to extract يَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (he longs for) that has the same pattern but whose HEAD’s root S-b-b was not in our list of modality roots.

Among the new extracted AM-MWEs are the expressions َيَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (he longs for) that occurred only once in the corpus. We used the pattern to extract َيَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (he longs for) that has the same pattern but whose HEAD’s root S-b-b was not in our list of modality roots.

Among the new extracted AM-MWEs are the expressions َيَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (he longs for) that occurred only once in the corpus. We used the same pattern to extract َيَذَا ﻣِنَ ﻣِنَ َيَلَوْنَ An (he longs for) that has the same pattern but whose HEAD’s root S-b-b was not in our list of modality roots.

This bootstrapping stage results in 358 AM-MWEs. They are inspected during manual verification.

5 Manual Verification and Final Results

We manually verify the best clusters, the bootstrapped AM-MWEs and the constructed patterns before including them in the final lexicon and repository to guarantee accuracy. Besides, we manually add modality senses to the lexicon entries. We also manually complete the morphological paradigms of the morphologically productive AM-MWEs. That is, if we only have the bigram يَرَغِبُ ﻓِﻲُ (he longs for) conjugated for the third singular masculine person, we manually add the rest of the conjugations.

The final lexicon is represented in XML and is organized by modality senses and then roots within each sense. The lexicon comprises 10,664 entries. The XML fields describe: the Arabic string, the size of the AM-MWE, the corpus frequency and the pattern ID. The pattern ID is the link between the lexicon and the repository because it maps each lexicon entry to its lexical, POS pattern in the repository.

<table>
<thead>
<tr>
<th>Roots</th>
<th>Senses</th>
<th>Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-m-l</td>
<td>Epistemic</td>
<td>4233</td>
</tr>
<tr>
<td>A-k-d</td>
<td>Evidential</td>
<td>811</td>
</tr>
<tr>
<td>r-g-b</td>
<td>Obligative</td>
<td>748</td>
</tr>
<tr>
<td>S-E-r</td>
<td>Permissive</td>
<td>755</td>
</tr>
<tr>
<td>H-s-s</td>
<td>Commissive</td>
<td>111</td>
</tr>
<tr>
<td>q-n-E</td>
<td>Abilitive</td>
<td>676</td>
</tr>
<tr>
<td>E-q-d</td>
<td>Volitive</td>
<td>330</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigrams</td>
</tr>
<tr>
<td>Trigrams</td>
</tr>
<tr>
<td>Quadrigrams</td>
</tr>
</tbody>
</table>

Table 5: Statistics for the AM-MWE lexicon for the top 7 roots and the distributions of modality senses and AM-MWE sizes

If a lexicon entry is manually added, the tag MANUAL is used for the corpus frequency field. Table 5 gives more statistics about the lexicon in terms of modality senses, AM-MWE sizes and the top 7 frequent modality roots.

The XML repository is given in the three POS tagsets supported by MADAMIRA. The XML fields describe: the pattern’s ID, the POS of the head and the pattern itself with the HEADs and GAPS marked. Appendices A and B give snapshots of the lexicon and the repository in Buckwalter's POS tagset.

6 Conclusion and Outlook

We described the unsupervised construction of a lexicon and a repository of variation patterns for AM-MWEs to boost their automatic identification and extraction. In addition to the creation of novel resources, our research gives insights about the morphological, syntactic and lexical properties of such expressions. We also propose an evaluation methodology that accounts for the productive insertion patterns of AM-MWEs and their morphological variations.

For future work, we will work on larger AM-MWEs to cover insertion patterns that we could
not cover in this paper. We will experiment with different association measures such as point-wise mutual information. We will also try different clustering algorithms.

Acknowledgement

This work was supported by grant NPRP 09-410-1-069 of the Qatar National Research Fund. We would also like to thank the four anonymous reviewers for their constructive comments.

References

Dragos Stefan Munteanu and DanielMarcu. 2007. ISI Arabic-English Automatically Extracted Parallel Text, Linguistic Data Consortium, Philadelphia
Michael Stubbs. 2007. An Example of Frequent English Phraseology: Distributions, Structures and
Appendix A: A snapshot of the XML lexicon

<lexicon name="AM-MWE Lexicon v1.0">
  <modality sense="abilitive">
    <head root="q-d-r">
      <am-mwe string=" ﻷد fostar" len="2" freq="283" patternID="23" />
      <am-mwe string=" لديه القترة على" len="3" freq="7" patternID="45" />
      ...
    </head>
  </modality>

  <modality sense="epistemic">
    <head root="g-l-b">
      <am-mwe string=" ﻓﻲ اﻟﻐﺎﻟﺐ" len="2" freq="122" patternID="15" />
      <am-mwe string=" ﺑﺒPEAR ﻷد fostar" len="2" freq="70" patternID="10" />
      ...
    </head>
  </modality>

  <head root="H-w-l">
    <am-mwe string=" ﻷد fostar" len="4" freq="38" patternID="50" />
    ...
  </head>
</lexicon>

Appendix B: A snapshot of the XML repository

<repository name="AM-MWE Variation Patterns v1.0">
  <tagset name="Buckwalter" pos-tagger="MADAMIRA v1.0">
    <pattern ID="10" head-pos="*+H4+*" pos="(HEAD)+(An/SUB_CONJ)" />
    ...
    <pattern ID="15" head-pos="DET+NOUN+*" pos="(ly/NOUN)+(PRON/GAP)+(HEAD)+(EIY/PPREP)" />
    ...
    <pattern ID="50" head-pos="DET+NOUN+*" pos="(mn/PPREP)+(HEAD)+(ADV/GAP)+*(An/SUB_CONJ)" />
    ...
  </tagset>
</repository>