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].

1 Introduction

Arabic Modal Multiword Expressions (AM-MWEs) are complex constructions that convey modality senses. We define seven modality senses, based on Palmer's (2001) cross-lingual typology, which are (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 حتما ولابد *HtmA wlAbd* (must), لعل وعسى *lEl wEsY* (maybe) and فيما يبدو *fymA ybdw* (seemingly).

Type 2 covers morphologically productive expressions such as يرغب في *yrgb fy* (he wants to) and واثق من *wAvq mn* (sure about). They inflect

[1] Both resources are available at http://www.rania-alsabbagh.com/am-mwe.html

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 114–123,
Gothenburg, Sweden, 26-27 April 2014. ©2014 Association for Computational Linguistics

AM-MWEs		Unigram Synonym(s)		English Gloss
Arabic	**Transliteration**	**Arabic**	**Transliteration**	
عقدت العزم على	*Eqdt AlEzm ElY*	عزمت - نويت	*Ezmt - nwyt*	I intended (to)
في امكاني ان	*fy AmkAny An*	يمكنني	*ymknny*	I can/I have the ability to
لدي اعتقاد بان	*ldy AEtqAd bAn*	اعتقد	*AEtqd*	I think
هناك احتمال بان	*hnAk AHtmAl bAn*	يُحْتَمَل	*yuHotamal*	possibly/there is a possibility that

Table 1: Example AM-MWEs and their unigram synonyms

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-modals (i.e. modals 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 الضروري *AlDrwry* (necessary) in من الضروري ان *mn AlDrwry An* (it is necessary to) with الممكن *Almmkn* (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 من الضروري جدا ان *mn AlDrwry jdA An* (it is very necessary to). Furthermore, we can have an inserted indirect object as in من الضروري للمصريين ان *mn AlDrwry llmSryyn An* (it is necessary for Egyptians to).

Type 5 includes morphologically, lexically and syntactically productive expressions like لدي يقين ان *ldy yqyn An* (I have faith that). Morphologically, the object pronoun in لدي *ldy* (I have) inflects for person, gender and number. Syntactically, the head noun can be modified by adjectives as in لدي يقين راسخ ان *ldy yqyn rAsx An* (I have a strong faith that). Lexically, the meaning of the expression relies on the head noun يقين *yqyn* (faith) which is replaceable for other modality-based nouns such as لدي نية ان *ldy 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 من الضروري ان *mn AlDrwry An* (literally: from the necessary to; gloss: it is necessary to), the preposition من *mn* (from) cannot be replaced by any other preposition. In هناك احتمال بان *hnAk AHtmAl bAn* (there is a possibility that), the relative adverbial هناك *hnAk* (there is) cannot be replaced by another relative adverbial such as هنا *hnA* (there is). In يحدوني الامل في ان *yHdwny AlAml fy An* (hope derives me to), the head is the noun الامل *AlAml* (the hope). Therefore, the lexical verb يحدوني *yHdwny* (drives me) cannot be replaced by other synonymous verbs such as يقودني *yqwdqny* (leads me) or يدفعني *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.

Corpus	Token #	Types #	Description
Ajdir	113774517	2217557	a monolingual newswire corpus of Modern Standard Arabic
LDC ISI	28880558	532443	an LDC parallel Arabic-English corpus (Munteanu & Marcu 2007)
YADAC	6328248	457361	a dialectal Arabic corpus of Weblogs and tweets (Al-Sabbagh & Girju 2012)
Tashkeel	6149726	358950	a vowelized corpus of Classical and Modern Standard Arabic books
Total	**41472307**	**3566311**	

Table 2: Statistics for the extraction corpora

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[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-linguistic information such as document and segment IDs, section headers, dates and sources, as well as English data are removed. (3) Punctuation marks are separated from words. (4) Words in Roman letters are removed. (5) Orthographical normalization is done so that all *alef*-letter variations are normalized to *A*, the elongation letter (_) and word lengthening are removed. (6) Finally, the corpus is tokenized and Part-of-Speech (POS) tagged by MADAMIRA (Pasha et a. 2014); the latest version of state-of-the-art Arabic tokenizers and POS taggers.

4.1.2 Extraction Set-up and Results

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*)***n***(\w*) to identify words such as الممكن *Alm**mkn*** (the possible), اتمكن *At**mkn*** (I manage) and بامكاني *bA**mk**A**n**y* (I can) which convey modality.

This RegExp-based procedure can result in noise. For instance, the aforementioned RegExp also returns the word الامريكان *AlA**m**ry**k**A**n*** (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.

[2] Ajdir: 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[3]. 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: $\mathrm{LLR} = 2\ \sum_{ij} \mathrm{O}_{ij} \log \frac{\mathrm{O}_{ij}}{\mathrm{E}_{ij}}$

Table 3 shows the unique type counts of the extracted ngrams. The extracted ngrams include both modal and non-modal MWEs. For instance, both من الممكن لنا ان *mn Almmkn lnA An* (it is possible for us to) and في اقرب وقت ممكن *fy Aqrb wqt mmkn* (as soon as possible) are extracted as valid quadrigrams. Both have the word ممكن *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 Almmkn lnA An* is an AM-MWE according to our definition in Section 2, but *fy Aqrb 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.

Ngram size	**Unique Types**
Bigrams	86645
Trigrams	43397
Quadrigrams	25634
Total	**96031**

Table 3: Statistics for the extracted MWEs

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 ض-ر-ر ***D-r-r*** include ضروري *__Dr__w__r__y* (necessary), بالضرورة *bAl__Dr__w__r__p* (necessarily), and يضطر *y__D__T__r__* (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 ان *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

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

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 $ngram_i$ is the probability of $ngram_i$ to occur in a sentence-initial position. It is computed as the frequency of $ngram_i$ occurring at a sentence-initial position normalized by the total number n of $ngram_i$ in the corpus.

Similarly, POS features are continuous. For instance, the probability that $ngram_i$ is followed by a deverbal noun is the frequency of its POS_{+1} tagged as a deverbal noun normalized by the total number n of $ngram_i$ 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[4]. 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 F_1-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 يتمكن من *ytmkn mn* (he manages to), يعجز عن *yEjz En* (he fails to) and يحلم ب *yHlm be* (he longs for), (2) prepositional phrases such as **من الممكن** *mn Almmkn* (it is possible that) and **في** الحقيقة *fy AlHqyqp* (actually), (3) nominal phrases such as املي هو *Amly hw* (my hope is to) and (4) AM-MWEs subcategorizing for complementizers such as يصرح بان *ySrH bAn* (he declares that) and يعرف ان *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 يفشل في ان *yf$l fy An* (he fails to) and prepositional phrases like من المستحيل ان *mn AlmstHyl An* (it is impossible to) and عندي ايمان بان *Endy AymAn bAn* (I have faith that).

The positive **QuadSet** includes verb phrases such as يحدوني الامل **في ان** *yHdwny AlAml fy An* (hope drives me to) and prepositional phrases such as من غير المقبول ان *mn gyr Almqbwl An* (it is unacceptable to).

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 quadrigram 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 quadrigram 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 يتمكن من *ytmkn mn* (he manages to) include its exact match and the morphological alternations of **اتمكن من** ***A**tmkn mn* (I manage to) and **نتمكن من** ***n**tmkn mn* (we manage to), among others. 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 instances such as يتمكن الرئيس من ***ytmkn*** *Alt}ys* ***mn*** (the president manages to) or **نتمكن نحن من** ***ntmkn*** *nHn* ***mn*** (we manages to), we consider them as true positives for the bigram *ytmkn mn* as long as the trigram has the two defining words of the bigram, namely the verb *ytmkn* in any of its conjugations and the preposition *mn*.

The same bigram - *ytmkn mn* - can have two gaps after the head verb *ytmkn* as in يتمكن الرئيس **المصري من** ***ytmkn*** *Alr}ys AlmSry* ***mn*** (the Egyptian president manages to). For that reason, we consider the best quadrigram cluster. If we find ***ytmkn*** ((pro)noun gap) ((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 quadrigrams.

False positives for the bigram *ytmkn mn* would be the bigrams يتمكن الرئيس ***ytmkn*** *Alr}ys* (the president manages) and الرئيس من *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 quadrigram best clusters to evaluate trigram AM-MWEs. We also tolerate morphological productivity.

For instance, the trigram عندنا ايمان بان *End**nA*** *AymAn bAn* (we have faith that) conjugated for the first person plural is a true positive for the gold set trigram عندي ايمان بان *En**dy** AymAn bAn* (I have faith that), 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 عند الناس ايمان بان ***End*** *AlnAs* ***AymAn bAn*** (people have faith that). The second can be an adjective after the head noun *AymAn*. Thus we can have عندي ايمان مطلق بان ***Endy AymAn*** *mTlq* ***bAn*** (I have a strong faith that).

Consequently, in the output of the quadrigram 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 عند الناس ايمان مطلق بان ***End*** *AlnAs* ***AymAn*** *mTlq* ***bAn*** (people have a strong faith that). Since we do not extract pentagrams, we consider chunks such as عند الناس ايمان ***End*** *AlnAs AymAn* (people have faith) and ايمان مطلق بان *AymAn mTlq* ***bAn*** (strong faith that) as false positive trigrams. This is because the former misses the complementizer **بان** *bAn* (in that), and the latter misses the first preposition **عند** *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, يحدونا الامل **في ان** *yHdw**nA** AlAml fy An* (hope drives us to) is considered as a true positive for يحدوني الامل في ان *yHdw**ny** AlAml 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.

	Precision	**Recall**	**F_1**
Bigrams	0.663	0.776	0.715
Trigrams	0.811	0.756	0.783
Quadrigrams	0.857	0.717	0.780

Table 4: Clustering evaluation results

Long dependencies are a source of errors at the recall level. Clustering could not capture such instances as **صرح** الرئيس المصري حسني مبارك**ب** ***SrH*** *Alr}ys 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 يتوق الى *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*)***t***(\w*)***w***(\w*)***q***(\w*) that was supposed to extract the volitive يتوق *ytwq* (he longs) did not return any instances for the intended modal but rather instances for يتوقف ***ytwqf*** (he stops) which interestingly subcategorizes for a preposition and a complementizer as in يتوقف عن ان *ytwqf En An* (literally: stops from to). This subcategorization frame is the same for modals such as يعجز عن ان *yEjz En An* (literally: unable from to). Consequently, يتوقف عن ان *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 @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 @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 HEAD tag to the head words (i.e. words whose roots were used for extraction)
- We add a GAP tag for adverbs, pronouns and other gap fillers to explicitly mark gap locations

An example pattern for the root ط -م - ح *T-m-H* (wish) is ((HEAD/*IV*) + (*AlY*/PREP) + (*An*/SUB_CONJ)) which reads as follows: a trigram AM-MWE whose head is a verb in any conjugation followed by the preposition *AlY* (to) and the subordinate conjunction *An* (that; to). Another pattern that results from the aforementioned steps for the same root of *T-m-H* is ((HEAD/*IV*) + (ADV/GAP) + (*AlY*/PREP) + (An/SUB_CONJ)). It means that an adverb can be inserted in-between the HEAD and the preposition *AlY* (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*) + (*AlY*/PREP) + (*An*/SUB_CONJ)) to extract the third person feminine plural conjugation of the root *T-m-H* in the trigram يطمحن الى ان *yTmHn AlY An* (they wish for) that occurred only once in the corpus. We used the same pattern to extract يصبو الى ان *ySbw AlY 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 من الواضح ان *mn AlmwADH An* (it is clear that) and من الطبيعي ان *mn AlTbyEy An* (it is normal that) that share the same pattern with من الممكن ان *mn Almmkn An* (it is possible that). We decide to consider those expressions as AM-MWEs although they are not epistemic in the conventional sense. That is, they do not evaluate the truth value of their clause-based propositions, but rather presuppose the proposition as true, and express the speakers' sentiment towards it.

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 يرغب في *yrgb fy* (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.

Roots		**Senses**		**Sizes**	
A-m-l	710	Epistemic	4233	Bigrams	4806
A-k-d	693	Evidential	811	Trigrams	3244
r-g-b	396	Obligative	748	Quadrigrams	2614
$-E-r	378	Permissive	755		
H-s-s	370	Commissive	111		
q-n-E	312	Abilitive	676		
E-q-d	293	Volitive	3330		
Total: 10,664					

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

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another Dialectal Arabic Corpus. *Proc. of LREC'12*, Istanbul, Turkey, May 23-25 2012

Rania Al-Sabbagh, Jana Diesner and Roxana Girju. 2013. Using the Semantic-Syntactic Interface for Reliable Arabic Modality Annotation. *Proc. of IJCNLP'13*, Nagoya, Japan, October 14-18 2013

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genbith. 2010. Automatic Extraction of Arabic Multiword Expressions. *Proc. of the Workshop on MWE 2010*, Beijing, August 2010

Elsaid Badawi, M.G. Carter and Adrian Gully. 2004. *Modern Written Arabic: A Comprehensive Grammar*. UK: MPG Books Ltd

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistic Package. *Proc. of CiCling'03*, Mexico City, USA

Ibrahim Bounhas and Yahya Slimani. 2009. A Hybrid Approach for Arabic Multi-Word Term Extraction. *Proceedings of NLP-KE 2009*, Dalian, China, September 24-27 2009

Kristen E. Brustad. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian and Kuwaiti Dialects*. Georgetown Uni. Press, Washington DC, USA

Tim Buckwalter. 2002. Arabic Morphological Analyzer. Technical Report, Linguistic Data Consortium, Philadelphia

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. *Proc*. of *the 51st ACL*, , Sofia, Bulgaria, August 4-9 2013

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed Belief Annotation and Tagging. *Proc. of the 3rd LAW Workshop, ACL-IJCNLP'09*, pp. 68-73, Singapore

Carla Parra Escartín, Gyri Smørdal Losnegaard, Gunn Inger Lyse Samdal and Pedro Patiño García. 2013. Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes. *Proc. of eLex 2013*, pages 338-357, Tallinn, Estonia, October 17-19 2013

Abdelati Hawwari, Kfir Bar and Mona Diab. 2012. Building an Arabic Multiword Expressions Repository. *Proc. of the 50th ACL*, pages 24-29, Jeju, Republic of Korea, July 12 2012

Andrew Lampert, Robert Dale and Cecile Paris. 2010, Detecting Emails Containing Requests for Action. *Proc. of the 2010 ACL*, pages 984-992, Los Angeles, California, June 2010

F. Mitchell and S. A. Al-Hassan. 1994. *Modality, Mood and Aspect in Spoken Arabic with Special Reference to Egypt and the Levant.* London and NY: Kegan Paul International

Ola Moshref. 2012. Corpus Study of Tense, Aspect, and Modality in Diglossic Speech in Cairene Arabic. PhD Thesis. University of Illinois at Urbana-Champaign

Dragos Stefan Munteanu and Daniel Marcu. 2007. ISI Arabic-English Automatically Extracted Parallel Text, Linguistic Data Consortium, Philadelphia

Malvin Nissim and Andrea Zaninello. 2013. A Repository of Variation Patterns for Multiword Expressions. *Proc. of the 9th Workshop of MWE*, pp. 101-105, Atlanta, Georgia, June 13-14 2013

Frank R. Palmer. 2001. *Mood and Modality*. 2nd Edition. Cambridge University Press, Cambridge, UK

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran and Irena Koprinska. 2013. *Automatically Detecting and Attributing Indirect Quotations*. *Proc. of the 2013 EMNLP*, pages. 989-1000, Washington, USA, October 18-21 2013

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proc. of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 26-31 2014

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written Dialog and Social Power: Manifestations of Different Types of Power in Dialog Behavior. *Proceedings of the 6th IJCNLP*, pp. 216-224, Nagoya, Japan, October 14-18 2013

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of CiCling 2002*, pages 1-15, Mexico City, Mexico

Michael Stubbs. 2007. An Example of Frequent English Phraseology: Distributions, Structures and

Functions. *Language and Computers: Corpus Linguistics 25 Years on*, pages 89-105, (17)

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="قادر على" len="2" freq="283" patternID="23"> </am-mwe>
        <am-mwe string="لديه القدرة على" len="3" freq="7" patternID="45"> </am-mwe>
        ...
      </head>
    </modality>
    <modality sense="epistemic">
      <head root="g-l-b">
        <am-mwe string="في الغالب" len="2" freq="122" patternID="15"> </am-mwe>
        ...
      </head>
      <head root="H-w-l">
        <am-mwe string="يستحيل ان" len="2" freq="70" patternID="10"> </am-mwe>
        ...
      </head>
      <head root="n-Z-r">
        <am-mwe string="من المنتظر ايضا ان " len="4" freq="38" patternID="50"> </am-mwe>
        ...
      </head>
    </modality>
</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="*+IV+*" pos="(HEAD)+ (An/SUB_CONJ)"></pattern>
      ...
      <pattern ID="15" head-pos="DET+NOUN+*" pos="(fy/PREP)+(HEAD)"></pattern>
      ...
      <pattern ID="23" head-pos="ADJ+*" pos="(HEAD)+(ElY/PREP)"> </pattern>
      ...
      <pattern ID="45" head-pos="DET+NOUN+*" pos="(lyd/NOUN)+(PRON*/GAP)*+(HEAD)+(ElY/PREP)">
      </pattern>
      ...
      <pattern ID="50" head-pos="DET+NOUN+*" pos="(mn/PREP)+(HEAD)+(ADV/GAP)*+(An/SUB_CONJ)">
      </pattern>
      ....
   </tagset>
</repository>
```