Inflectional and Morphological Variation of Arabic Multi-word Expressions

Dhekra Najar(✉), Slim Mesfar, and Henda Ben Ghezala

RIADI, University of Manouba, Manouba, Tunisia
dhekra.najar@gmail.com, mesfarslim@yahoo.fr, hhbg.hhbg@gmail.com

Abstract. CompounDic is an Arabic MWEs dictionary that lists many entries, divided into more than 20 domains. It lists only MWEs in their base form. With regard to syntactic and morphological flexibility, the lexicon covers 2 types of MWEs: Fixed MWEs (no variation allowed) and semi-fixed MWEs (variation in their structural pattern). Arabic presents distinctive features to deal with MWEs processing. A lot of possible derivations are possible (plural or dual forms, multiple irregular plurals). In addition, we need to process agglutination forms. In this paper, we will study the structural variability of semi-fixed multiword expressions in Arabic language in order to recognize the morphological and inflectional variations. We will adopt a recognition approach based on the use of a cascade of local grammars.

The recognition system is based on NooJ's local grammars as well as an Arabic MWEs dictionary covering more than 20 domains. The inflectional and derivational rules, which concern semi-fixed MWEs, use some specific morphological operators that will be described as well. Finally, we present new results showing the experimentation scores of morpho-lexical coverage enhancement.

Keywords: Multi-word expressions · Natural language processing · NooJ · Arabic language · Compound words variation

1 Introduction

A Multi-Word Expressions (MWEs) are groups that work together as units to express a specific meaning. They can be formed by combining two or more words together. Generally, lexical and morphological analyzers are not able to recognize multiword expressions unless they are listed in internal resources. Automatic analyzers usually process MWE as separated terms. As a result, semantics is lost because generally the meaning of the MWE is different from the meanings of its components.

Most multi-word expressions allow certain types of variability on their components. This problem has to be taken into account for their description to be able to recognize them in texts as well as their potential variations.

The identification of MWEs is essential for any natural language processing based on lexical information. Therefore, recognizing only the limited MWEs that are usually listed in computational lexicon is not enough. The morphological and inflectional variability of MWEs and their lexical particularities need to be described in the computational lexicon

L. Barone et al. (Eds.): NooJ 2016, CCIS 667, pp. 37–47, 2016.
DOI: 10.1007/978-3-319-55002-2_4

in order to be able to recognize the full range of their occurrences in texts. The rest of the paper is organized as follows: Sect. 2 describes expressions topology as well as their structural variability and presents the MWE"s lexicon CompounDic. The proposed approach is discussed in Sect. 3. Section 4 shows the experimental results. Section 5 summarizes the results of this work and draws conclusions.

2 Multi-word Expressions

2.1 Arabic MWE's Variability Types

Based on previous works, we identify three types of variability of MWEs: fixed, semi-fixed and syntactically flexible.

- Fixed MWEs are considered as a list of words with spaces and with no morphological variation allowed. This category contains unambiguous compound expressions such us (Middle East, الشرق الأوسط) and frozen sentences such us pragmatically fixed expressions (مَدَى الحياة, forever) and proverbs.
- Semi-fixed expressions allow variations including graphical variants, which are the graphic alternations between the letters (ى ,ي) and the letters (ه ,ة), as the following illustrates. As well, many morphological variants can effect semi-fixed expressions. Specifically, we mention variations that express person, number, tense, gender, and the definite article that is carried out by the fixed morpheme (ال, Al) (Fig. 1).

Fig. 1. Example of inflectional variants of an entry.

- While MWEs that are syntactically flexible allow new external elements (components) to intervene between the MWE components (Fig. 2).

Arabic words are characterized by their complex structure. In comparison with Semitic languages, Arabic language presents distinctive features, namely the vocalization that causes a lexical ambiguity in texts. Also, Arabic is an agglutinative language (the prefix (definite article (the, ال), prepositions (for, لِ) and (with, بِ), conjunctions (and, و), suffixes (her, ـه)).

The Arabic language has a complex MWEs structure (up to 5 units) and a lot of possible variations and derivations (dual forms, multiple irregular plurals… ect). The recognition of all potential inflected and agglutinated forms attached to each entry needs

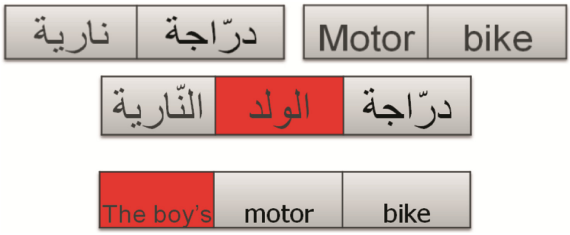

Fig. 2. Example of syntactically flexible of an entry.

a special tokenization that depends on their linguistic specificities. However, we used to make some specific tools to be able to deal with the specificities of the Arabic language.

Arabic presents distinctive features to deal with MWEs processing. A lot of particular variations are possible:

- Agglutinated forms;
- Inflectional variations: (Gender and number: plural or dual forms, multiple irregular plurals).
- Morphological Variations: (Definite article, Personal agglutinated pronouns, Agglutinated conjunctions and prepositions).

2.2 CompounDic

In previous work, we have semi-automatically built CompounDic (Najar et al. 2015), an Arabic 2 units MWEs thematic lexicon. For this purpose, we have taken advantage of NooJ's[1] linguistic engine strength in order to create this large coverage terminological MWEs dictionary for Modern Standard Arabic language CompounDic. NooJ is a linguistic development environment that allows formalizing complex linguistic phenomena such as compound words generation, processing as well as analysis.

However in Nooj "simple words and multi-words units are processed in a unified way: they are stored in the same dictionaries, their inflectional and derivational morphology is formalized with the same tools and their annotations are undistinguishable from those of simple words" (Silberztein 2005).

CompounDic contains 36960 entries classified into more than 20 semantic domains. It covers the category of fixed expressions except proverbs and semi-fixed expressions as well as the different types of MWEs such as expressions that are traditionally classified as idioms, prepositional verbs, collocations, and so on. In this lexicon, we didn't deal with flexible expressions.

All the entries of CompounDic are manually set in the base form: "indefinite singular form". Then, all the listed MWEs are voweled manually so that NooJ would be able to recognize unvoweled, semi-voweled as well as fully voweled MWEs. The manual vocalization is an extremely important step since it allows to vowel entries depending

[1] http://www.nooj4nlp.net/.

on their semantic information since we can find a word that has different way of vocalization and different meanings. This helps reducing linguistic ambiguities in Arabic texts.

The final manual step is classifying the MWEs according to 2 criteria: the grammatical composition (N1 N2), (N1 ADJ, and so on).

In fact, the Arabic MWE can be a combination of different forms: a verb, a noun, an adjective and a particle. Most of MWEs are composed of one or more nouns (N), adjectives (ADJ), adverbs (ADV) or simple named entities. We provide the syntactic phrase structure composition of our Arabic MWEs, giving each entry of our lexical resource its component elements (noun + noun, noun + adjective, verb + preposition + noun…).

We manually extract a list of about 15 patterns of MWEs compositions classified into 4 basic categories (Table 1):

Table 1. Patterns of MWEs compositions

Type	Structure
	ADJ_N
Descriptive compound	ADJ_prep
	ADJ_prep_N
	ADJ1_ADJ2
	N1_N2
	N1_prepN2
Compound nouns	N_ADJ
	N_prep
Negation	Neg_N
	neg_V
Prepositional nouns	prep_N
Compound Verbs	V_N
	V_prep
	V_prepN
	V1_V2

The entries of CompounDic are classified into more than 20 domains as shown in Table 2.

Every entry in CompounDic is stored with information about its structure, number of units and domain. To give a simple example from the technical domain in our lexicon:

العِدَام إتِّزان, **N + Structure = N1_N2 + CMPD + Units = 2 + Domain = Technical.**

As it was said, fixed MWEs always occur in exactly the same structure and can be easily recognized by a lexicon. However, most MWEs allow different types of

Table 2. Number of entries in CompounDic per domain

Domain	Entries	Domain	Entries
sportive	1434	economical	2222
financial	2268	media	508
agriculture	2401	educational	1526
political	3251	religious	1645
press	44	UN	1582
military	2323	Touristic	2251
legal	2191	Computer	2058
psycological	2068	weather	329
social	2572	transport	2356
Industry	582	engineering	1277
administrative	578	technical	2076
	Total		36960

modifications. In Arabic language, we can reach an average of 33 possible variations to each MWE entry. Arabic presents distinctive features to deal with MWEs processing such as plural or dual forms, multiple irregular plurals and agglutination forms. With this in mind, we still have a lot of possible variations to recognize from CompounDic lexicon.

3 Approach

In order to improve Natural Language Processing system performances, it is important to identify MWEs in texts since it helps to disambiguate semantic and lexical content. Generally speaking, we have 2 potential solutions to recognize CompounDic entries variations:

- **Generation method:** focuses on inflectional and derivational descriptions that are manually implemented for each MWE entry. This method is not efficient due to the exponential complexity that can cause and the time that take to manually implement descriptions.
- **Recognition method**: focuses on lexical grammars that recognize the MWE's variations. This method uses local grammars to recognize the related forms of CompounDIC entries without generating them. Usually, the result of the recognition method is precise. Furthermore, it processes agglutinated forms. However, we will be faced to heavy linguistic analysis since NooJ will check the lexical constraints for each digram.

In view of this, we propose to use the recognition method with based-rules local grammars in order to automatically recognize the inflectional and morphological variations from CompounDic entries using NooJ's linguistic engine. We are going to add

some enhancement to this method in order to avoid heavy linguistic analysis especially while processing big corpus.

To sum up, our system will be able to:

- Recognize the morphological and inflectional variations of Arabic MWEs.
- Annotate MWEs in text with their distributional (Domain = Financial…) and syntactic information (Noun + Noun, Noun + Adj…).
- Get a better semantic representation.
- Reduce the lexical and syntactic ambiguity.

4 Grammar

We are going to use NooJ's linguistic engine to implement a local grammar describing the structural variability of Arabic MWEs. This grammar will be able to recognize all the morphological and inflectional variants of CompounDic entries, namely:

- Gender (female, male);
- Number (dual, plural);
- Definite article: the fixed agglutinated morpheme (ال, Al);
- Personal agglutinated pronouns;
- Agglutinated conjunctions and prepositions (for, لِ), (with, بِ), (and, وَ).

As noted earlier, the enhancement of the recognition method is important to avoid heavy linguistic analysis. For this reason, we are going to focus the analysis on the units that are attested to be a part of a MWE.

- Step 1: extract all the units of our CompounDIC.
- Step 2: add to the extracted units in El_DicAr the distributional information (+CmpElem).

To do this, we have developed a program to enrich El-Dicar[2]. It allowed us to add semi automatically about 2000 unknowns (technical words) and automatically 7000 distributional information (+CmpElem). We are still working on the enrichment of El-DicAr dictionary.

We illustrate this semi-automatic enrichment program by Fig. 3.

[2] Electronic Dictionary for Arabic "El-DicAr" resources (Mesfar et al. 2008), developed using NooJ's linguistic engine.

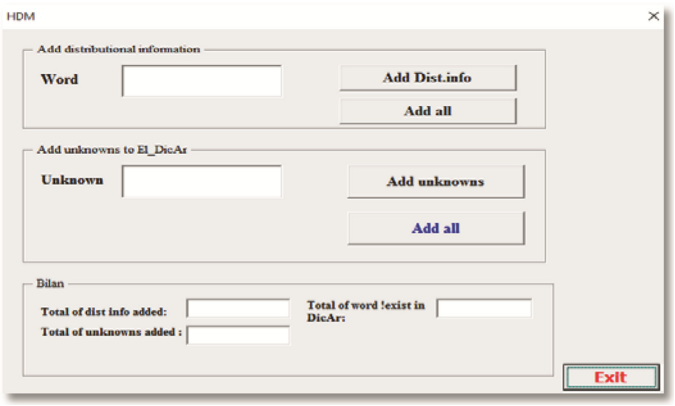

Fig. 3. Enrichment program platform.

Our local grammars are implemented based on the 17 patterns of MWEs compositions that we have extracted as shown previously. As we can see in Fig. 4 we have the grammar structure that shows all the MWEs structures and the main grammar of our system.

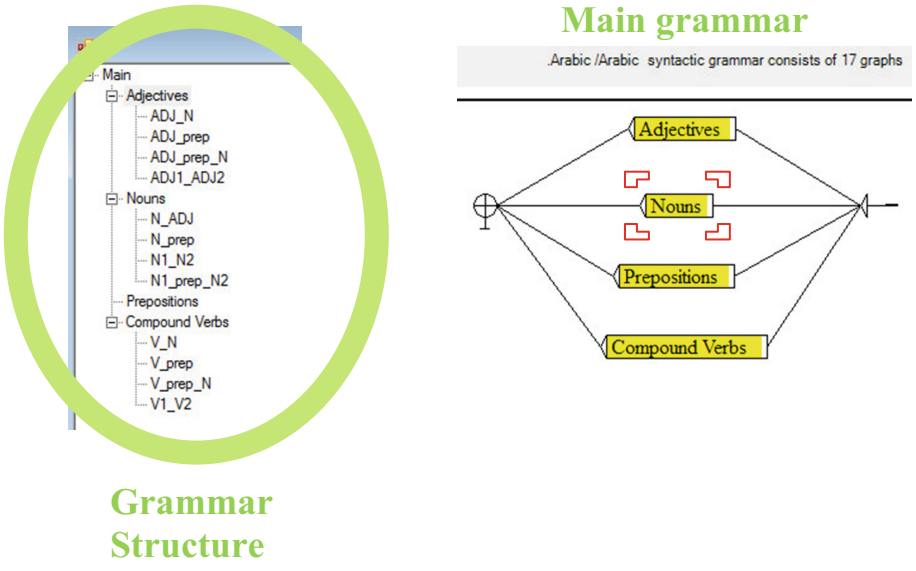

Fig. 4. Local grammar structures and the main sub graph.

With the distributional information +CmpElem, the linguistic analysis of our grammar will be limited on the units that are attested to be a part of a MWE. To do so, we are going to use distributional information +CmpElem in the grammar to identify

MWEs components. To demonstrate this, we give an illustration of a sub graph of MWE structure composed of 2 units: **NOUN_ADJ**.

As shown in the Fig. 5, N and ADJ are 2 Variables to save each digram element to use them in a lexical constraint. The sub graph, as seen in Fig. 5, indicates the constraints below:

1. *$N_# #$ADJ_ = : N + CMPD + Structure = N_ADJ*[3]
   - Concatenate the 2 lemmas.
   - Compare N and ADJ values (in base form) with CompounDIC entries.
   - Restrict the comparison only to the defined structure.
   - Annotate the recognized MWEs variations with Semantic description (+CMPD + Domain + Structure).
   - Recognize agglutinated forms (prepositions: < PREP >, prefix: < PREF >, pronoun: < PRON >).
2. *$N_ $ADJ_, N$1S>*
   - $N_: Represents the lemma of the lexical unit stored in $N variable
   - N$1S: inherits the semantic information (Domain) from the recognized MWE to annotate the matching sequence.

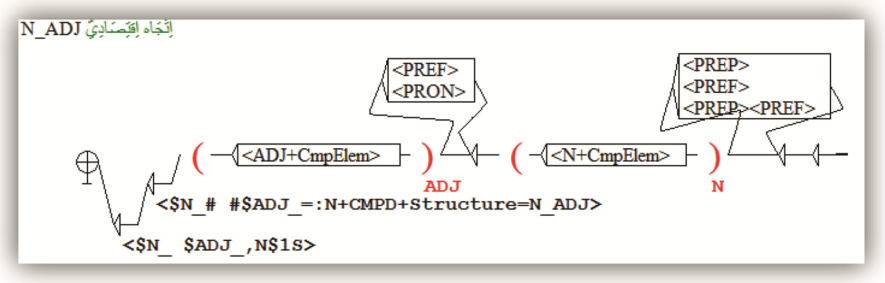

Fig. 5. MWEs variations local grammar NOUN_ADJ

Demonstrating this, the grammar process a text when it finds two or simple words with the distributional information (+CmpdElem): it will put each word in a variable $Var_ tracked by "_" to set them to their base form (indefinite Singular form). All the stored consecutive variables will be concatenated < $Var1_ $Var2_ > to get the same multi-word expression but in the base form. Then, the grammar will try to find a similar entry of the MWE in our lexicon using the first constraint (1).

Once the MWE is found, it will be recognized and considered as a variation of an existing MWE in CompounDic lexicon. The grammar allows inheriting the semantic information (Domain) from the recognized MWE.

However, we have a particular case of entries containing agglutinated prepositions (V_prepN, N1_prepN2, ADJ_prepN) as shown in the sub graph below. It's not possible for our grammar to recognize agglutinated MWE elements. So, we have made some changes in the constraints of sub graphs of MWEs structures with agglutinates elements.

[3] NooJ's syntactic, inflectional and semantic categories are detailed in Annex.

To be specific, we give the example of the prepositional structure NOUN1_prepNOUN2 (Fig. 6).

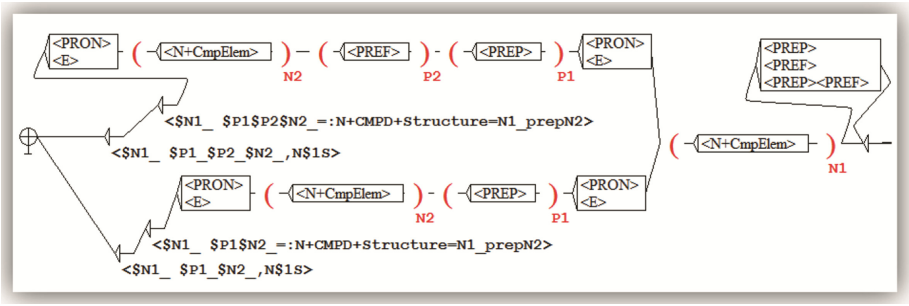

Fig. 6. MWEs variations local grammar NOUN1_prepNOUN2

The same thing with the first example except:

1. $N1_ $P1$P2$N2_ = : N + CMPD + Structure = N1_prepN2
   Concatenate the 2 lemmas including the prepositions (without modifying the form of the prepositions). Check in CompounDIC entries. If it exists in our lexicon then it will be considered as a variation of MWE.
2. *$N1_ $P1_ $P2_$N2_, N$1S*
   – N$1S: inherits the semantic information (Domain) from the recognized MWE to annotate the matching sequence.

5 Results and Discussion

To test the lexical recognition of our grammar, we launched the linguistic analysis of our test corpus. We presented preliminary experiments on a corpus containing 870 heterogeneous articles from Internet. We reported high quality result.

The table above presents the recall and precision obtained by testing the grammar on the test corpus. The results, as seen in Table 3, indicate that we have reached high quality results of recognition. Our results in term of precision (0.97 of precision) are better than other existing approaches. We presented preliminary experiments on a Concordance:

Table 3. Results

Precision	Recall
0.97	0.88

We believe that this automatic method ameliorated the precision of the results by recognizing all MWEs forms in the text (Fig. 7).

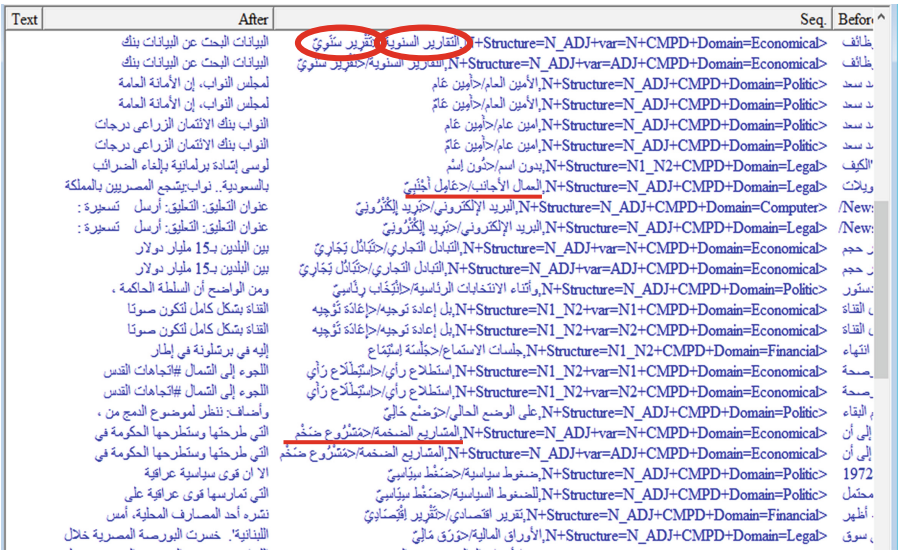

Fig. 7. Concordance

Illustrating the concordance, our grammar recognized expressions such as:

- (والمشاريع الضخمة **huge projects**): definite expression in the plural.

The base form of this expression in our lexicon is (مشروع ضخم, huge project).

Several obstacles make the recognition of Arabic MWE's variations really complicated such as high inflectional nature, morphological ambiguity related to some agglutinated forms, variant sources of ambiguity (unvoweled texts…) and dual forms for pronouns and verbs. These specificities of Arabic language represent the most challenging problems for Arabic NLP researchers.

More specifically, the silence in our grammar is due to some problems in CompounDic lexicon such as:

- False vocalization of words such as (misplaced vowels);
- Common typographical errors such as confusion between Alif and Hamza or the substitution of (ة ,ه) and (ى ,ي) at the end of the word;
- Lexical ambiguity of some agglutinated forms;
- Lack of entries in our lexicon.

6 Conclusion

MWEs are combinations of single terms expressing various meaning compared to the combination of single word's meanings. This paper focuses on recognizing multi-word expressions inflectional and morphological variations in Arabic corpus. Our research

has shown that rule-based approaches are more efficient in recognizing the entire multi-word expressions variations, especially morphological variations. We believe that this automatic method has improved the precision of the results.

Further research is needed to better understand the topology of MWEs in different languages.

7 Annex

NooJ's syntactic categories:

Syntactic codes	
<ADJ>	Adjective
<V>	Verb
<N>	Noun
<ADV>	Adverb
<CONJ>	Conjunction
<PREP>	Preposition
<PREF>	Prefix
<PRON>	Pronoun
<REL>	Relative pronoun
<PART>	Particle
<E>	Empty caracter
<P>	Ponctuation
Inflectional codes	
<s>	Singular
<p>	Plurial
<m>	Male
<f>	Female
Semantic codes	
<CmpdElem>	Component of a MWE

References

Najar, D., Mesfar, S., Ghezela, H.B.: A large terminological dictionary of Arabic compound words. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.) Automatic Processing of Natural-Language Electronic Texts with NooJ, pp. 16–28. Springer, Cham (2015)

Mesfar, S.: Analyse Morpho-syntaxique Automatique et Reconnaissance Des Entités Nommées En Arabe Standard. Thesis, Graduate School - Languages, Space, Time, Societies. Paris, France (2008)

Silberztein, M.: Nooj's dictionaries. In: Vetulani, Z. (ed.): Proceedings of the 2nd Language and Technology Conference. Wydawnictvo Poznańskie Sp. z o.o., Poznan (2005)