

A Large Terminological Dictionary of Arabic Compound Words

Dhekra Najar, Slim Mesfar^(✉), and Henda Ben Ghezela

RIADI, University of Manouba, Manouba, Tunisia

Dhekra.najar@gmail.com, mesfarslim@yahoo.fr,

hhbg.hhbg@gmail.com

Abstract. NooJ is a linguistic development environment that allows formalizing complex linguistic phenomena such as compound words generation, processing as well as analysis. We will take advantage of NooJ's linguistic engine strength in order to create a new large coverage terminological compound word's dictionary for Modern Standard Arabic language. Classifying and annotating Arabic compound words would have a major impact on the disambiguation of applications working with Arabic texts. The diverse analyzers, based on morphological aspect, are not able to recognize multiword expressions. Morphological analyzers usually separate compound expressions into single terms. Therefore recognizing the entire compound words is essential to preserve the semantic of texts and to provide a crucial resource for a better analysis and understanding of Arabic language.

Our work is composed of three sections. First, we will deal with a literature review on Arabic compound expression's categories which aims to dress a detailed topology. The structural variability of multiword expressions in Arabic language will be studied in order to measure the degree of morphological, lexical and grammatical flexibility of multiword expressions. Then, we will discuss the electronic thematic dictionary of compound Arabic expressions and give detailed description of our methodology and guidelines.

Keywords: Compound expressions · Natural language processing · NooJ · Arabic language

1 Introduction

Natural Language Processing (NLP) is known as the ability of systems to process natural languages. There are some steps that are considered standard in NLP. In order to process texts, usually computers divide textual documents into sets of units, known as terms. Within, there are many technical terms that often take form of Multiword expressions (MWE), particularly in specialized articles such as biomedical domain, economical domain, ect. Recent developments in the field of Natural Language Processing have shown the need for recognizing MWEs in a text in order to avoid context ambiguities. For instance, (Nakagawa and Mori 2002) show that more than 85 % of domain-specific terms are multi-word terms. These MWE are combinations of single terms expressing different meaning compared to basic term's meaning. Of course

humans can easily identify multi-word expressions while processing natural languages, but for digital systems, MWEs will be semantically seen and analyzed as separate units. As a result, semantics is lost because generally the meaning of the MWE is different from the meanings of the components.

Table 1. Example of MWE

Compound word	Components	Comments
كأس العالم world cup	كأس: cup العالم: world	Disambiguating meanings

As it can be seen from Table 1, the lemma “world” has ten different senses, in Wordnet 1.6, and the lemma “cup” has four different senses (While the MWE “world cup” has only one sense). Morphological analyzers generally separate MWEs into single terms and this would have a negative effect on the accuracy of all applications working with textual documents. Therefore recognizing the entire compound expressions is essential to constitute a better representation of text semantic content than single word terms. The main purpose of the paper is to build a large multi-genre coverage MWE’s dictionary for Modern Standard Arabic (MSA) using NooJ’s linguistic engine. The remainder of the paper is organized as follows: Sect. 2 describes a literature review on Arabic MWE’s categories and topology. Also, the structural variability of multiword expressions in Arabic language will be studied in order to measure the degree of morphological, lexical and grammatical flexibility. The proposed thematic dictionary of MWEs is discussed in Sect. 3. Section 4 shows the experimental results. Section 5 summarizes the results of this work and draws conclusions.

2 Related Work and Typology of Arabic MWE

A multi-word expression is a consecutive sequence of at least two simple forms and blocks of separators (Silberztein 1993). There are three main approaches for extracting MWEs:

- Linguistic approach that makes use of lexicons and language rules such as morphological or syntactic information;
- Statistical approach that uses a set of standard statistical scores to estimate the degree of association between the words and their frequency in texts;
- Hybrid approach that combines the statistical and linguistic approach.

The majority of the latest MWE extraction systems have adopted the hybrid approach especially on Arabic. Statistical methods could not be applied straight since Arabic is a highly agglutinative language. The development of a terminology extraction system for Arabic language requires linguistic specifications of terms and should be

associated to its specific rules. However, to our best knowledge, very few publications can be found in the literature that discusses this issue in Arabic language.

For example, (Bounhas and Slimani 2009) have followed a hybrid method to extract compound terms. First, they detect compound noun boundaries and identify sequences that are likely containing compound nouns. Then, they use syntactic rules to handle MWEs. These rules are based on linguistic information: a morphological analyzer and a POS tagger. (Attia 2008) presented a pure linguistic approach for handling Arabic MWEs. It is based on a lexicon of MWEs constructed manually. Then the system tries to identify other variations using a morphological analyzer.

Based on an analysis of the literature (Attia 2008), MWEs cover expressions that are traditionally classified as:

- idioms (e.g. the cold war), (الحرب الباردة);
- prepositional verbs (e.g. to come near to), (اقترب من);
- verbs with particles (e.g. to give up);
- compound nouns (e.g. a book cover), (غلاف الكتاب);
- collocations (e.g. howling of a wolf), (عواء الذئب);
- Conjunctive expression (assistance and cooperation), (التعاون والتآزر);

With regard to syntactic and morphological flexibility, MWEs are classified into three types: fixed, semi-fixed and syntactically flexible. Fixed compound nouns are entered in the lexicon as a list of words with spaces with no morphological variation allowed. This category contain unambiguous compound expressions such as (Middle East, الشرق الأوسط) and frozen sentences such as pragmatically fixed expressions (مَدَى الحياة, forever) and proverbs. The variations that can effect semi-fixed expressions include graphical variants, which are the graphic alternations between the letters (ي, ى) and the letters (ة, ه), as the following illustrates (Fig. 1).

مستشار	تسويقي
Marketing	consultant

مستشار	تسويقي
Marketing	consultant

Fig. 1. Example of graphical variants.

In fact, graphical variants depend on the author origins. They are processed using some low priority morphological grammars (Mesfar 2008).

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). Figure 2 shows an example of inflectional variants of an entry in Arabic language.

While MWEs that are Syntactically Flexible allow new external elements (components) to intervene between the MWE components.

In the next section, we will define the linguistic specifications of Arabic MWEs and present the implementation steps of our dictionary.

Definite article		Number and gender	
تفصيلي	خبر	التفصيلي	الأخبار
detailed	news	The detailed	the news
Plural form		تفصيلية	أخبار
Dual form		تفصيليتين	خبرين

Fig. 2. Example of inflectional variants.

3 Arabic MWEs Dictionary

3.1 Linguistic Specifications of Arabic MWE

Arabic words are characterized by its 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, ل) and (with, ب), conjunctions (and, و), suffixes (her, ه)).

Arabic language has a complex MWEs structure (up to 5 units) and a lot of Arabic language has a complex MWEs structure (up to 5 units) and a lot of possible variations and derivations (dual forms, multiple irregular plurals...). The recognition of all potential inflected and agglutinated forms attached to each entry needs 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 (Mesfar 2010).

3.2 Specifications of Our Lexicon

Many researches on MWE recognition in literatures have especially focused on biomedical domain. In our approach, we will organize the multi-words entries, composed of 2, 3, 4 units and more, of our lexicon into 20 semantic fields.

Our lexicon is covering: Fixed expressions except proverbs; semi-fixed expressions and their Inflectional variants that will be processed using morphological grammars. These grammars will be recognizing all the morphological variants of the related forms:

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

Our lexicon covers the different types of MWEs such as expressions that are traditionally classified as idioms, prepositional verbs, collocations, etc.

Table 2. The semantic fields of our Lexicon

Religious terms	Educational terms	Medical terms	Journalistic terms	Politic terms
Social terms	Technical terms	Administrative terms	Financial terms	Economic terms
Transport terms	Weather terms	Sportive terms	Restaurant and touristic terms	Engineering terms
Agricultural terms	Computer sciences terms	Military terms	Press terms	Industrial terms
Psychological terms	Legal terms	Media terms	Organisation terms	BioMedical terms

3.3 Proposed Approach

NooJ linguistic engine is based on large coverage dictionaries and grammars. It uses Finite State Transducers (FSTs) to parse text corpora made up of hundreds of text files in real time and associate each recognized entry with its related information, such as morpho-syntactic information (POS - Part Of Speech, Gender, Number, etc.), syntactic and semantic information (e.g. transitive, Human, etc.). NooJ is a well known linguistic environment that is already used to formalize more than 20 languages.

The recognition process consists on identification of lexical entities using dictionaries and grammars, and the transformation of grammars into transducers.

NooJ¹ 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)

We will take advantage of NooJ’s linguistic engine strength in order to create a new large coverage terminological MWEs dictionary for Modern Standard Arabic language.

Firstly, a lexicon of MWEs is collected manually and associated with a set of semantic information. Data were gathered from various online Arabic linguistic web-sites. This morphological lexicon contains lexical entries divided into more than 20 domains. Most of these entries belong to scientific and technical terminology. The rest

¹ <http://www.nooj4nlp.net/>.

حُرِّيَّة	Freedom	حُرِّيَّة التِّجَارَةِ	Freedom of trade
Part of MWE		حُرِّيَّة الإِجْتِمَاع	Freedom of assembly

Fig. 3. Example of a part of MWE

of entries are extracted semi automatically from Arabic corpus using NooJ's linguistic engine. A list of thematic relevant terms that frequently occur as part of an MWE in a specialized text is built. For example and as shown in Fig. 3, we state the term "freedom" in a legal corpus.

Using local grammars, these terms are then tracked by other units in a concordance and the output is added to our MWEs lexicon. Secondly, noisy data will be manually eliminated or rectified in the lexicon:

- Common typographical errors such as confusion between Alif and Hamza or the substitution of (ة, ة) and (ي, ي) at the end of the word;
- The false writing of Hamza;
- The addition or omission of a character in a word;
- The lack of white between two terms

Thirdly, all the entries of the lexicon are set in the base form: "indefinite singular form" in order to automatically generate the flexional and derivational forms using NooJ's local grammars that we will implement.

Then, all the listed MWEs were voweled manually so that NooJ would be able to recognize unvoweled, semi-voweled as well as fully voweled MWEs. In some cases of Arabic words, we can find a word that has different way of vocalization and different meanings. So, the manual vocalization is an extremely important step since it allows us to vowel entries depending on their semantic information. 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 the number of elements (1, 2, 3, 4...).

Table 3. Patterns of MWEs compositions

Grammatical category	2 units Patterns	3 units Patterns
Prothetic compound (مركَّب اضافي)	N1_N2	N1_N2_N3 N1_ADJ_N2 N1_ADJ_prepN2
Descriptive compound (مركَّب نعتي)	N1_ADJ	ADJ1_ADJ2_ADJ3 N1_ADJ1_ADJ2 N1_ADJ1_prepADJ2
Compound verb (فعل مركَّب)	V_N V_prep	
Attributive compound (مركَّب شبه اسنادي)	N1_prepN2 N_prepADJ	
Adjective noun	ADJ_N	

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 (only 2 and 3 units 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 13 patterns of MWEs compositions.

Moreover, each lexical compound entry of our lexicon is associated with a set of semantic (Semantic information where we cover semantic fields) and distributional information (see Table 2). Organizing our specialized lexical entries in semantic field format brings many practical benefits; one of those is to allow classifying textual documents by category and translating texts by themes.

Entries in our lexicon are structured as follow:

MWEEntry, N+CMPPD+Struct=GrammaticalComposition+Length=n+FIELDname

Example: **حق شرعي**, N+CMPPD+Struct=N_ADJ+Taille=2+Juridique

3.4 The Structural Variability of Our MWEs Lexicon

For the following parts of this work, we will use the Electronic Dictionary for Arabic “El-DicAr” resources (Mesfar 2008) as the basis of our local grammars for MWEs variations recognition using NooJ’s morphological analyzer. Our approach is essentially based on a manually constructed lexicon of MWEs. Then the system tries to identify other variations of the MWEs which concern semi-fixed compound words. A semi-fixed multiword expression is a frequent combination of two words or more, characterized by high degree of morphological and syntactic flexibility.

So, how can we recognize all the variations and the agglutinated forms of the lexicon’s semi-fixed MWEs?

		Def. Singular form		Ind. Prep. form	
وجهي	تعبير	الوجهي	التعبير	وجهي	بتعبير
facial	expression	The facial	the expression	facial	With expression
Def. Prep. form		Ind. Singular pron. form		Ind. plural form	
الوجهي	بالتعبير	الوجهي	تعبيره	وجهية	تعبير
The facial	With the expression	The facial	His expression	facial	expressions
Def. plural form		Ind. plural pron. form		Ind. Dual form	
الوجهية	التعبير	الوجهية	تعبيره	وجهين	تعبيرين
The facials	the expressions	The facial	His expressions	Two facials	Two expressions

Fig. 4. Variations and the agglutinated forms of an entry

To illustrate, we present the different variations and derivations of the psychological MWE (تعبير وجهي, facial expression) as shown in Fig. 4.

With the new NooJ's V5 there is no more need to create inflectional paradigms for compound words since NooJ can reuse the inflectional paradigms for existing simple words. For example, the MWE (طبيب شرعي, forensic pathologist) has as Inflexional paradigm:

طبيب, N+Job+FLX=Atibbea26c
 شرعي, ADJ+FLX=AdjDesc1
 طبيب شرعي, N+CMPD+Struct=N_ADJ+Taille=2+FLX=AdjDesc1<P>Atibbea26c

And it recognizes the variations below:

Fem.Sing	طَبِيبَةٌ شَرْعِيَّةٌ	Forensic pathologist
Masc.Pl	أَطْبَاءُ شَرْعِيَّونَ	Forensic pathologist
Fem.Dual	طَبِيبَتَيْنِ شَرْعِيَّتَيْنِ	Two Forensic pathologists
Masc.Dual	طَبِيبَيْنِ شَرْعِيَّيْنِ	Two Forensic pathologists

As we notice, the example respect gender and number because it's a human noun. Unfortunately, this solution could not be used for the whole Arabic MWEs flexions. Arabic MWEs do not always respect gender and number agreement when generating the associated forms like in Latin languages, especially for irregular plural forms. To illustrate, we give the example of (Fig. 5):

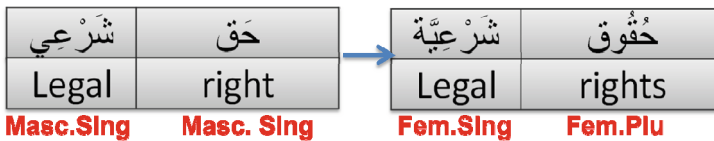


Fig. 5. Example of irregular MWE variation

In comparison with the singular form, the MWE do not respect the gender and number agreement while generating the plural. Hence, identifying the morphological flexibility and variations (graphical variants, inflectional variants...) they may have in the plural and dual forms should be either through the generation method (1) or the recognition method (2);

Solution 1: Generation Method. It is based on building new appropriate flexional and derivational descriptions that are manually implemented for each MWE entry to

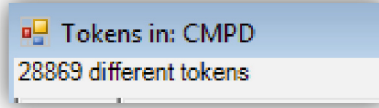


Fig. 6. Number of units in our lexicon (+CmpdElem)

generate the list of associated forms. This first approach obtains an exponential complexity due to the multiple derived forms (Exponential complexity when checking gender and number agreements manually). Also, it is time consuming to build the flexional and derivational descriptions.

Solution 2: Recognition Method. It is based on local grammars that recognize the MWE's variations and related forms without generating them. With this solution, we can process agglutinated forms as well. However, a number of limitations need to be considered. Particularly, the heavy linguistic analysis since NooJ will check the lexical constraints for each bigram, trigram....

Solution 2 bis (Extended Version): We opt for implementing a system based on the recognition method associated with some enhancements to reduce heavy linguistic analysis. In order to restrict the analysis into the units who are attested to be a part of a MWE, all the units (simple words) of our MWEs lexicon were separately extracted and annotated in El_DicAr with the distributional information (+CmpdElem). As shown in the figure, we have extracted about 28870 different units from our lexicon (Fig. 6).

Then, we develop a local grammar containing all the identified grammatical patterns (see Table 3). This grammar would be able to recognize the duals and plural forms (regular and irregular forms) of a MWE as well as its different agglutinated forms. For example, we present the grammar of the 2 units pattern N1_ADJ (Fig. 7):

If the grammar, while processing texts, finds two or more consecutive 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 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 annotation:

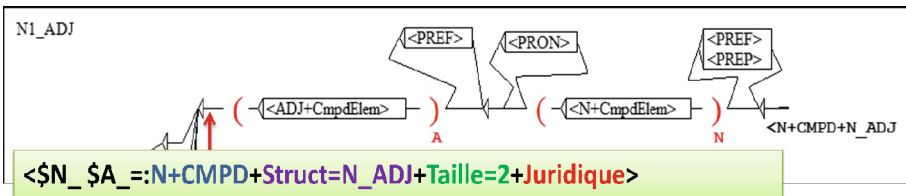


Fig. 7. MWEs variations local grammar

$\langle \$Var1_ \$Var2_ \dots =: N + CMPD + Struct = Structural\ Composition + lenght = n + FIELD \rangle$.

Once the MWE is found, it will be recognized and considered as a variation of an existing MWE in our lexicon.

We have built a large compound word's dictionary covering multiple domains for Modern Standard Arabic using NooJ's linguistic engine that is able to:

- Recognize all the potential inflected forms.
- Annotate MWEs in texts.
- Improve the lexical coverage of the Electronic Arabic dictionary El_DicAr.
- Get a better semantic representation.
- Reduce the lexical ambiguity.

4 Evaluation and Results

Several obstacles make the analysis of Arabic MWEs so complicated: the high inflectional nature (uses internal patterns for its grammatical processes), the agglutinated form of pronouns and prepositions, the variant sources of ambiguity (unvowled texts...), the dual forms for pronouns and verbs. These specificities of Arabic language represent the most challenging problems for Arabic NLP researchers. Compared to the big amount of available resources and MWEs lexicons in Latin languages, particularly English, the Arabic language is still immature.

We have collected about 63500 MWEs (base form) associated to 20 different fields, as shown in Table 4. We believe that these rates are high especially considering all the variations (inflectional and agglutinated forms) that can have each entry of the lexicon.

We note that 2 units MWEs represent more than half of all entries in the lexicon (66 %) followed by 3 units MWEs (21.8 %). The rest of entries are considered as fixed expressions and do not undergo MWEs variations recognition.

To test the lexical coverage of our dictionary, we launch the linguistic analysis of our corpora. We present preliminary experiments on a corpus containing 150 heterogeneous journalistic articles (Fig. 8).

The table above presents the recall and precision obtained by testing the coverage of our lexicon on the test corpus. The results, as seen in Table 5, 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.

These are several possible explanations for the low rate of the recall:

- False vocalization of words such as (misplaced vowels);
- Common typographical errors such as confusion between Alif and Hamza or the substitution of errors and (ة, ا) at the end of the word;
- Lexical ambiguity of some agglutinated forms in test corpus;
- Some delimitation problems related to some incomplete MWEs in our lexicon;
- Lack of entries in our dictionary.

Table 4. The lexicon’s MWEs entries

Semantic category	2 units	3 units	4+ units	Total
Economical	2253	960	563	3776
Media	540	184	103	827
Educational	1564	334	163	2061
Religious	1682	653	293	2628
Organization	1624	1083	992	3699
Touristic	2384	481	416	3281
Computer	2241	690	455	3386
Weather	339	145	69	553
Transport	2433	844	474	3751
Engineering	1230	439	116	1785
Technical	2129	655	347	3131
Biomedical	3888	1217	483	5588
Sportive	1447	619	326	2392
Financial	2393	879	553	3825
Agriculture	2411	750	106	3267
Political	3351	750	292	4393
Press	44	45	48	137
Military	2354	933	596	3883
Social	2628	812	384	3824
Psychological	2121	602	259	2982
Administrative	587	168	89	844
Total	41906	13886	7640	63432

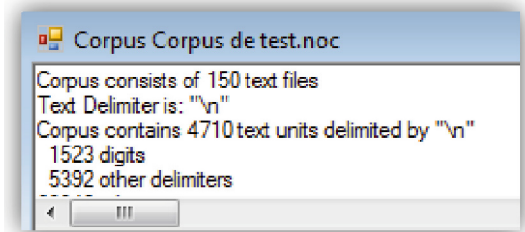


Fig. 8. Corpus of Test

Table 5. Results

Precision	Recall
0.97	0.88

5 Conclusion and Perspectives

To sum up, multi-word terms have a great importance since they constitute domain relevant candidate terms. The study has shown that recognizing the entire compound words is essential to preserve the semantic of texts and to provide a crucial resource for understanding Arabic language. We believe that this semi-automatic method has reduced linguistic ambiguities and has improved the precision of the results. Our results in term of precision are better than other existing approaches.

More research is needed to better understand the topology of MWEs in different languages. To improve the effectiveness of an information retrieval system, further research should be done to investigate the possibility to match our MWEs lexicon with a view to ontology construction. A further point is to identify semantic relations between the concepts of the linguistic ontology to get a better semantic analysis of texts.

Annex :

NooJ's syntactic categories:

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

References

- Nakagawa, H., Mori, T.: A simple but powerful automatic term extraction method. In: COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology, vol. 14, pp. 1–7. Association for Computational Linguistics (2002)
- Silberztein, M.: Les groupes nominaux productifs et les noms composés lexicalisés. In: *Linguisticae Investigationes XVII: 2*. John Benjamins B.V., Amsterdam (1993)
- Bounhas, I., Slimani, Y.: A hybrid approach for Arabic multi-word term extraction. In: International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009, pp. 1–8. IEEE (2009)
- Attia, M.: Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. Thèse de doctorat, University of Manchester (2008)
- Mesfar, S.: Analyse Morpho-syntaxique Automatique et Reconnaissance Des Entités Nommées En Arabe Standard. Thesis, Graduate School—Languages, Space, Time, Societies, Paris, France (2008)
- Mesfar, S.: Towards a cascade of morpho-syntactic tools for Arabic natural language processing. In: Gelbukh, A. (ed.) *CICLing 2010*. LNCS, vol. 6008, pp. 150–162. Springer, Heidelberg (2010)
- Silberztein, M.: NooJ’s dictionaries. In: The Proceedings of the 2nd Language and Technology Conference, Poznan (2005)
- Mesfar, S.: Analyse morpho-syntaxique et reconnaissance des entités nommées en arabe standard. Thèse, Université de franche-comté, France (2008)

Automatic Processing of Natural-Language Electronic
Texts with Nooj

9th International Conference, Nooj 2015, Minsk,
Belarus, June 11-13, 2015, Revised Selected Papers
Okrut, T.; Hetsevich, Y.; Silberztein, M.; Stanislavenka,
H. (Eds.)

2016, XII, 227 p. 135 illus., Softcover

ISBN: 978-3-319-42470-5