A Framework for the Classification and Annotation of Multiword Expressions in Dialectal Arabic

Abdelati Hawwari, Mohammed Attia, Mona Diab
Department of Computer Science
The George Washington University
{Abhawwari,mohattia,mtdiab}@gwu.edu

Abstract

In this paper we describe a framework for classifying and annotating Egyptian Arabic Multiword Expressions (EMWE) in a specialized computational lexical resource. The framework intends to encompass comprehensive linguistic information for each MWE including: a. phonological and orthographic information; b. POS tags; c. structural information for the phrase structure of the expression; d. lexicographic classification; e. semantic classification covering semantic fields and semantic relations; f. degree of idiomaticity where we adopt a three-level rating scale; g. pragmatic information in the form of usage labels; h. Modern Standard Arabic equivalents and English translations, thereby rendering our resource a three-way – Egyptian Arabic, Modern Standard Arabic and English – repository for MWEs.

1 Introduction

Multiword expressions (MWEs) comprise a wide range of diverse, arbitrary and yet linguistically related phenomena that share the characteristic of crossing word boundaries (Sag et al., 2002). MWEs are computationally challenging because the exact interpretation of an MWE is not directly obtained from its component parts. MWEs are intrinsically single units on the deep conceptual and semantic levels, but on the surface (lexical and syntactic) levels they are expressed as multiple units. MWEs vary in their syntactic category, morphological behavior, and degree of semantic opaqueness. MWEs are pervasively present in natural texts, which makes it imperative to tackle them explicitly if we aspire to make large-scale, linguistically-motivated, and precise processing of a human language.

Integrating MWEs in NLP applications has evidently and consistently shown to improve the performance in tasks such as Information Retrieval (Acosta et al. 2011; da Silva and Souza, 2012), Text Mining (SanJuan and Ibekwe-SanJuan, 2006), Syntactic Parsing (Eryiğit et al., 2011; Nivre and Nilsson, 2004; Attia, 2006; Korkontzelos and Manandhar, 2010), Machine Translation (Deksne, 2008; Carpuat and Diab, 2010; Ghoneim and Diab 2013; Bouamor et al., 2011), Question Answering, and Named-Entity extraction (Bu et al., 2011).

In the current work, we propose guidelines for detailed linguistic annotation of an MWE lexicon for dialectal (Egyptian) Arabic that covers, among other types, expressions that are traditionally classified as idioms (e.g. على الريق EalaY Alriyq[1] 'on an empty stomach'), prepositional verbs (e.g. توكل على tawak~al EalaY 'rely on'), compound nouns (e.g. إشارة مرور <i$Arap muruwr 'traffic light'), and collocations (e.g. أخد دش >axad du$~ 'to take a shower').

Creating a repository of annotated MWEs that is focused on dialects is essential for computational linguistics research as it provides a crucial resource that is conducive to better analysis and understanding of the user-generated content rife in the social media (such as Facebook, Twitter, blogs, and forums). Moreover, it helps in understanding he correspondences between different languages and their representation of the semantic space. We hope that the multilingual data in this repository will lead to a significant enhancement in the processing of comparable and parallel corpora. We believe that our proposed framework will contribute to the sustainability of

[1] In this paper, we use the Buckwalter Transliteration Scheme for rendering Romanized Arabic as described in www.qamus.com.

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 48–56,
October 25, 2014, Doha, Qatar.

MWE research in general, and provide a blue print for research on MWEs in dialects, informal vernaculars, as well as morphologically rich languages.

MWE are not only interesting from an NLP perspective but also from a linguistic perspective, as MWE can help in understanding the link between lexicon, syntax and semantics. Until now, this is hampered by the lack of comprehensive resources for MWEs with fine-grained classification on different dimensions related to semantic roles and syntactic functions. Arabic comprises numerous divergent dialects, and having an annotated MWE lexical resource in dialects and Modern Standard Arabic (MSA) will allow for studying transformation, change and development in this language.

From a theoretical linguistic point of view, our work will be interesting particularly in studies related to Diglossia. Diglossia (Walters, 1996) is where two languages or dialects exist side by side within a community, where typically one is used in formal contexts while the other is used in informal communications and interactions. Studying the MWE space for dialects and MSA as a continuum will lead to deeper insights into variations as we note intersection and overlap between the two. In many instances, we see that MSA MWEs and their dialectal equivalents are not necessarily shared as they occupy complementary linguistic spaces. However, the nature of this complementarity and its cultural and social implications will need more exploration and investigation, which will be possible once a complete resource becomes available.

In the current work, we give detailed description of our methodology and guidelines for annotating phonological, morphological, syntactic, semantic and pragmatic information of an Egyptian Multiword Expressions (EMWE) lexical resource. Our annotation scheme covers the following areas.

a) Phonological and orthographic information;

b) POS tag, based on the observation of how an MWE functions as a whole lexical unit;

c) Syntactic variability and structural composition;

d) Lexicographic types, which includes the classifications followed in the dictionary-writing domain (idioms, support verbs, compound nouns, etc.);

e) Semantic information, where we cover semantic fields and relations;

f) Idiomaticity Degree; we adopt a three level rating scale (Mel'čuk, 1998) to measure the degree of semantic opaqueness;

g) Degree of morphological, lexical and syntactic flexibility (Sag et al., 2002);

h) Pragmatic information, which includes adding usage labels to MWEs where applicable;

i) Translation, which includes the MSA and English equivalents, either as an MWE in MSA and English if available or as a paraphrase otherwise.

2 Previous Work

There are four main areas of research on MWEs: extraction from structured and unstructured data, construction of lexicons for specific languages, integration in NLP applications, and the construction of guidelines and best practices. A significant amount of research has focused on the identification and extraction of MWEs (Ramisch et al., 2010; Dubremetz and Nivre, 2014; Attia et al., 2010; Weller and Heid, 2010; Schneider et al., 2014). Description and specifications of MWE lexical resources have been presented for Japanese (Shudo et al. 2011), Italian (Zaninello and Nissim, 2010), Dutch (Grégoire, 2010; Odijk, 2013), and Modern Standard Arabic (Hawwari et al., 2012). Moreover, Calzolari et al. (2002) presented a project that attempted to introduce best practice recommendations for the treatment of MWE in mono- and multi-lingual computational lexicons that incorporate both syntactic and semantic information, but the limitation of their work is that they focus on only two types of MWEs, namely, support verbs and noun compounds.

Apart from Schneider et al. (2014), who focused on the language of the social web, none of these projects dealt with informal or dialectal languages, which are rampant in user-generated content (UGC). With the explosion of social media, the language of Web 2.0 is undergoing fundamental changes: English is no longer dominating the web, and UGC is outpacing professionally edited content.

UGC is re-shaping the way people are consuming and dealing with information, as the user is no longer a passive recipient, but has now turned into an active participant, and in many instances, a source or producer of information. Social media have empowered users to be more creative and interactive, and allowed them to

voice their opinions on events and products and exert powerful influence on the behavior and opinion of others. Yet, the current overflow of UGC poses significant challenges in data gathering, annotation and presentation.

3 MWE Taxonomy

Although the importance of the MWEs has been acknowledged by many researchers in the field of NLP as evident by the large number of research papers and dedicated workshops in the past decade, the theory of MWEs is still underdeveloped (Sag et al., 2002). There is critical need for studying MWEs both from the theoretical and practical point of views. MWEs have diverse categories, varying degrees of idiomaticity, different syntactic compositions, and different morphological, lexical and syntactic behavior, and dealing with them is complicated even further by the fact that there is no "watertight criteria" for distinguishing them them (Atkins and Rundell, 2008).

Moreover, there is no universally-agreed taxonomy of MWEs (Ramisch, 2012), and different researchers proposed different typology for this phenomena. Fillmore et al. (1988) proposed three types based on lexical and syntactic familiarity: a) unfamiliar pieces familiarly combined, b) familiar pieces unfamiliarly combined, and c) familiar pieces familiarly combined. Mel'čuk (1989), on the other hand, introduced three different classes: a) complete phraseme, b) semi-phraseme, c) and quasi-phraseme. Sag et al. introduced two classes: institutionalized phrases and lexicalized phrases, with lexicalized phrases subdivided into fixed, semi-fixed and syntactically flexible expressions. Ramisch (2012) introduced yet another set of classes: nominal, verbal and adverbial expressions.

From the lexicographic point of view, the legacy three-way division of MWEs proved to be too coarse-grained to cater for the needs of lexicographers who need to identify the large array of sub-types that fall under the umbrella of 'MWEs'. Atkins and Rundell (2008) emphasized the need for lexicographers to be able to recognize MWE types such as fixed phrases, transparent collocations, similes, catch phrases, proverbs, quotations, greetings, phatic expressions, compounds, phrasal verbs, and support verbs.

When we look deeply into the different classifications, we notice that each approach has looked at the phenomenon from a different angle, either focusing on its syntactic regularity, semantic and pragmatic properties, meaning compositionality, surface flexibility, POS (part of speech) category, or lexicographic relevance. What we propose is that it is not possible to come up with a hard and fast classification that cuts through all levels of representation. All afore-mentioned classifications are valid and can work parallel to each other, instead of substituting for each other. The assumption that we follow in this paper is that MWEs have different classifications at different levels of representation from the very deep level of semantics and pragmatics to the very shallow level of morphology and phonetics. The details of our annotation scheme are explained in the following section.

It is worth noting that in our current work, we move the focus away from edited text to the challenging and creative language found in UGC and by trying to close the language resource gap between edited and unedited text. We handle this gap by focusing on dialects, the language used in informal communications such as emails, chat rooms, and in social media in general. We cover the full range of MWEs (nominal, verbal, adverbial, adjectival and prepositional expressions) in Egyptian Arabic, covering 7,331 MWEs (collected from corpora and paper dictionaries).

4 Annotation of Linguistic Features in MWE

In this section, we provide a comprehensive specification of MWE types and the detailed linguistic information, including the phonological, orthographical, syntactic, semantic and pragmatic features.

4.1 Phonological

Each MWE is provided in full diacritization to indicate its common pronunciation in Cairene Arabic accent, such as عَلَى كَفّ عَفْرِيت *EalaY kaf~ Eaforiyt* 'at high risk', 'lit. on the palm of a demon'. We also list other phonological variants when available.

4.2 Orthography

Since dialects do not have a standard orthography, we follow the CODA style (Habash et al., 2012), which is a devised standard for conventionalizing the orthography of dialectal Arabic. CODA takes canonical forms and etymological

facts into consideration. For example, the Egyptian expression أخذ باله >axad bAluh 'to pay attention' is rendered in CODA as أخذ باله >axa* bAluh.

4.3 POS

At this level of annotation we consider the POS of the entire MWE when regarded as one unit from a functional perspective. We annotate each MWE with a POS tag from a predefined tagset. We define the POS tag based on the headword POS in the MWE. Our POS tagset includes verb, noun, adjective, adverb, interjection, proper noun, and preposition. The list of POS tags used along with examples is shown in Table 1.

	POS	Example
1	verb	جَرّ على الحِساب jar~ EalaY AlHisAb 'pay later'
2	noun	أكْل الْعِيْش >akol AlEay$ 'making ends meet' [lit. eating bread]
3	adjective	أشكال وألوان >a$okAl wa->alowAn 'various shapes and colors'
4	adverb	أخرة المتمة >axorip Al-matam~ap 'at the end'
5	interjection	يا ناس ياهوه yA nAs yAhuwh 'anybody there'
6	proper nouns	شجرة الدر $ajarip Aldur~ 'Shajar al-Durr'
7	preposition	بغض النظر عن bi-gaD~ AlnaZar Eano 'irrespective of'

Table 1: MWE Examples with their POS Tags

4.4 Syntactic Annotation

A syntactic variable is a slot that intervenes between the component parts of an MWE, without being itself a part of it, but fills a syntactic gap. Syntactic variables are added, when needed, to MWEs to represent the syntactic behavior of an MWE and they exemplify how the MWE interacts with other elements within its scope. We create a tagset of syntactic variables reflecting the argument structure of an MWE. Examples are shown in Table 2.

No	Syntactic Variable	Example
1	فلانُ somebody (masc_ nominative)	جسّ (فلانٌ) النبض jas~ (fulAn) AlnaboD 'somebody) tested the waters'
2	فلانةً somebody (fem_ accusative)	أكل (فلانة) بعينيه >akal (fulAnap) bi-Eaynayh 'he devoured (some woman/girl) with his eyes'
3	القومِ people (genitive)	دق بين (القوم) إسفين daq~ bayn (Alqawom) <isofiyn 'he drove a wedge between (some people)'
4	الأمرَ some matter (accusative)	حط (الأمر) في حسابه Hat~ (Al>amora) fiy HisAbihi 'he took (some matter) into consideration'
5	الشيءُ something (nominative)	(الشيء) متفصل عليه (Al$ayo') mitofaS~al Ealayh '(something) fits him perfectly'

Table 2: Syntactic variables and example usages

4.5 Lexicographic Annotation

In the dictionary market there are specialized dictionaries for idioms, phrasal verbs, proverbs and quotations. However, general domain dictionaries try to avoid the use of too technical terms in the description of MWEs and use for the sake of simplicity a general term like 'phrase' to denote them to users. Yet, in the meta language of the dictionary compiling profession, lexicographers make a more fine-grained distinction between the various types of MWEs. Our lexicographic classification of MWEs is adapted from Atkins and Rundell (2008) and includes the following tags. Examples are listed in Table 3.

1. Idiom: An idiom is an MWE whose meaning is fully or partially unpredictable from the meanings of its components (Nunberg et al., 1994);
2. Support verb, or 'light verbs', may be defined as semantically empty verbs, which share their arguments with a noun (Meyers et al., 2004);

3. Prepositional verb: These are verbs followed by prepositions with impact on the meaning;
4. Compound noun: A compound noun is a lexeme that consists of more than one noun;
5. Compound term: This is a technical compound noun used in a specific technical field;
6. Compound named entity: This is a multiword proper name;
7. Phatic expression: an expression that is intended for performing a social function (such as greeting or well-wishing) rather than conveying information;
8. Proverb: We consider proverbs as multiword expression if they are used as lexical units;
9. Quotation: We list only quotations that have gained currency in the language and have become familiar to the majority of the community.

	Classification	Example
1	Idiom	بيعمل من الحبة قبة biyiEomil min AlHab~ap qub~ap 'to make a mountain out of a molehill'
2	Support verb	أَخَد تار >axad tAr 'to take revenge'
3	Prepositional verb	ضحك عليه DiHik Ealayh 'to play a joke on' [lit. laugh on him]'
4	Compound noun	أبو قردان >abuw qirodAn 'Cattle egret'
5	Compound term	عرق النسا Eiroq AlnisA 'Sciatica'
6	Compound named entity	أَبُو الهول >abuw Alhuwl 'the Sphinx'
7	Phatic expression	أشوف وشك بخير >a$uwf wu$~ak bi-xayr 'see you later'
8	Quotation	يا مولاي كما خلقتني yA mawolAyA kamA xalaqotiniy 'penniless'
9	Proverb	العقل زينة AlEaqol ziynap 'wisdom is a blessing'

Table 3: Examples of Lexical Types

4.6 Structural Classification

We provide the syntactic phrase structure composition of the expressions, giving the MWE pattern or the POS of its component elements. The purpose is to show the normal productive syntactic patterns underlying the expressions. Table 4 shows the list of possible structural pattern in Egyptian MWEs.

	Structure	Example
1	adjective + conjunction + adjective	رَايق وَفايق rayiq wa-fayiq 'happy and relaxed'
2	adjective + noun	تنابلة السلطان tanaboliq Al-sulotAn 'couch potatoes' [lit. Sultan dependents]'
3	noun + noun	كِلِمِة حَق kilomiq Haq~ 'word of truth'
4	adjective + preposition + noun	غرقان لشوشته garoqAn li-uwotuh 'up to his ears'
5	adverb + noun	بين نارين bayn nArayn 'confused' [lit. between two fires]
6	adverb + verb	حَسْبَمَا اتَّفَقَ HasobamA Ait~afaq 'haphazardly' [lit. as happens]
7	noun + adjective	نفخة كدابه nafoxap kad~Abap 'false pride/arrogance' [lit. false blow]
8	verb + conjunction + verb	يِلِتّ وَيَعْجِنْ yilit~ wa-yiEojin 'to babble' [lit. knead and fold]
9	verb + verb	امشى انجرّ Aimo$iy Ainojar~ 'get moving/get out' [lit. walk and drag]
10	verb + preposition + noun	تَوكّل عَلَى الله tawak~al EalaY Allah 'rely on Allah/go away'
11	preposition + noun	على الطبطاب EalaY AlTabo-TAb 'effortlessly' [lit. on ease]
12	verb + noun	نفش ريشه nafa$ riy$uh 'show pride' [lit. stretched his feathers]
13	noun + verb	الله يرحمه! Allah yiroHamuh 'Allah have mercy on him'

Table 4: Examples Syntactic Classification

4.7 Semantic Fields

The entries in the current lexical resource are classified into semantic fields based on their semantic contents. The objective is to assign one semantic field tag for each MWE in the lexicon. Organizing Lexical data in semantic field format brings many theoretical and practical benefits, one of those is to allow the current lexical resource to function both as a lexicon and a thesau-

rus. In Table 5 we show a sample of our semantic field classification.

	Semantic Field	Example
1	Social Relation	سمن على عسل samon EalaY Easal 'getting on well' [lit. ghee on honey]
2	Oath and Emphasis	والله العظيم wa-Allah AlEaZiym 'I swear by Allah'
3	Occasions	يتربى في عزك yitrab~aY fiyEiz~ak 'congratulations on the new baby' [lit. may he grow up in your wealth]
4	Death	ربنا افتكره rab~inA Aifotakaruh 'he died' [lit. the Lord remembered him]
5	wishing and cursing	بَعْد الشّر baEod Al$ar~ 'God forbid' [lit. may the evil be far away]
6	trickery	لبّسه العمة lab~isuh AlEim~ap 'to hoodwink' [lit. put the turban on him]
7	Occultism	ضرب الرمل Darab Alramol 'to practice divination' [lit. to strike the sand]'

Table 5: Semantic fields

4.8 Semantic Relations

Aiming at presenting detailed lexical semantic information, we further classify our entries based on semantic relations like synonymy, antonymy and polysemy.

- Synonymy: MWE synonyms are grouped together; as the following expressions which all mean 'to practice divination' قرا الفنجان qarA AlfinojAn [lit. read the cup], ضرب الودع Darab AlwadaE [lit. hit the shells], قرَا الكف qarA Alkaf~ [lit. read the hand palm].
- Antonymy: MWE antonyms are two MWE having the opposite meaning to each other. For examples, إيده ناشفة <iyduh nA$ofap 'avaricious' [lit. his hand is dry] is the antonym of إيده مخرومة <iyduh maxoruwmap 'wasteful' [lit. his hand has a hole in it].
- Polysemy. This is when an MWE has more than one meaning. For example, إيده طويلة <iyduh Tawiylap [lit. his hand is long] can mean either a 'powerful person' or a 'thief'.

4.9 Idiomaticity Degree

Mel'čuk (1998) classified MWEs with regards to idiomaticity into three types: full phrasemes, quasi-phrasemes and semi-phrasemes.

- **Full phrasemes** are when the meaning of the expression does not match the meaning of the component words, such as وهلم جرا Wa-halum~ jar~A 'and so on'.
- **Quasi-phrasemes** are when the meaning of the expression matches the meaning of the component words in addition to an extra piece of meaning that is not directly derived from either components, such as مجلس الشعب majolis Al$aEob 'people's assembly'.
- **Semi-phrasemes** are when the meaning of the MWE is partially directly derived from one component and partially indirectly indicated by the other component, such as دراسات عليا dirAsAt EuloyA 'higher studies'.

4.10 Morpho-lexico-grammatical flexibility

A scale of three levels is used to measure the degree of morphological, lexical and grammatical flexibility of a MWE, adopted from Sag et al. (2002). The three levels are as follows:

- **Fixed MWE:** An MWE is considered as a fixed expression if it does not have any degree of syntactic, morphological or lexical flexibility, and its meaning cannot be predicted from its component elements, for example, سداح مداح sadAH madAH 'slapdash'.
- **Semi-Fixed MWE:** Semi-fixed expressions allow for a certain degree of morphological and lexical variation, but they are fixed in terms of the syntactic word order, for example, ماشية\ماشيين على حل شعرها\شعرهم mA$oyap/mA$oyiyn EalaY Hal~ $aEorahA/$aEoruhum [lit. living by letting down her/their hair] 'whore/whores' or 'loose women'.
- **Syntactically flexible MWE:** A syntactically flexible MWE is a frequent combination of two words or more, characterized by high degree of morphological and syntactic flexibility. Example, دش (فلان) إدى <id~aY

(fulAn) du$~ ‘to scold someone harshly’ [lit. give someone a shower].

4.11 Pragmatic Annotation (Usage Labels)

The reason we provide usage labels is inspired by the CALLHOME Egyptian Arabic corpus (Gadalla et al., 1997)), which is a collection of data gathered from spoken colloquial language. The usage labels present specifications on *who* uses an MWE and *how* it is used. The usage label tagset in our lexicon includes labels such *vulgar, youth, aggressive or taboo*, as exemplified in Table 6.

Who or how	Example
youth	يسوق الهبل في الجبل yisuwq Alhabal fiy Aljabal ‘to act foolishly’ [lit. to act madly in the mountain]’
women / girls	الشاطرة تغزل برجل حمار Al$ATrap tigozil birijol HumAr ‘make do with what you have’ [lit. a clever girl will knit with a donkey’s leg]’
Aggressive	أديك في وشك >ad~iyk fiy wi$~ak ‘I shall slap you in the face’

Table 6: Pragmatic annotation

5 Status of the current resource

The Egyptian MWE lexical resource at the current stage contains 7,331 entries, and work is still on going in the linguistic annotation of the dictionary. Table 7 presents the current annotation progress statistics regarding the various classifications and features.

	Feature	Completion
1	Diacritization	34.10%
2	Syntactic Variables	25.92%
3	MSA Equivalent	27.28%
4	POS	34.10%
5	Syntactic Classification	23.58%
6	English Equivalent	27.28%
7	Lexical Type	98.94%
8	Pragmatics Usage	4.09%
9	Synonymous	0.14%
10	Idiomaticity Degree	12.82%
11	Semantic-Field	2.29%

Table 7: Annotation work progress

6 Conclusion

We have described the annotation guidelines for a lexical database of MWE for dialectal Arabic. We provide descriptive specifications of MWE at the phonological, orthographical, syntactic and semantic levels. The main contribution of this paper is that it is the first description of a classification and annotation scheme of a lexical database for dialects, which can be extended for informal languages and with direct applicability on user-generated content.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency (DARPA) Contract No. HR0011-12-C-0014, BOLT program with subcontract from Raytheon BBN.

References

Acosta, Otavio Costa, Aline Villavicencio, Viviane P. Moreira. (2011) Identification and Treatment of Multiword Expressions applied to Information Retrieval. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 101–109, Portland, Oregon, USA, 23 June 2011.

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith. 2010. Automatic Extraction of Arabic Multiword Expressions. COLING 2010 Workshop on Multiword Expressions: from Theory to Applications. Beijing, China

Attia, Mohammed. (2006) Accommodating Multiword Expressions in an Arabic LFG Grammar. In T. Salakoski et al. (Eds.): Advances in Natural Language Processing. FinTAL 2006, *Lecture Notes in Computer Science*. Vol. 4139, pp. 87 - 98, 2006. Springer-Verlag Berlin Heidelberg.

Baldwin, T. (2005a). The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions* 19 (4), 398–414.

Baldwin, T. (2005b). Looking for prepositional verbs in corpus data. In Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Colchester, UK, pp. 115–126.

Baldwin, Timothy and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language

Processing, pages 267–292. CRC Press, Boca Raton, USA, 2nd edition.

Bannard, C. 2007. A Measure of Syntactic Flexibility for Automatically Identifying Multi Word Expressions in Corpora. Proceedings of A Broader Perspective on Multiword Expressions, Workshop at the ACL 2007 Conference: 1–8.

Benson, M. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23--35.

Bouamor, Dhouha, Nasredine Semmar and Pierre Zweigenbaum. (2011) Improved Statistical Machine Translation Using MultiWord Expressions. International Workshop on Using Linguistic Information for Hybrid Machine Translation LIHMT. Barcelona, November 2011

Bu, Fan, Xiao-Yan Zhu, and Ming Li. (2011) A New Multiword Expression Metric and Its Applications. In *Journal of Computer Science and Technology*. 26(1): 3-13, Jan. 2011.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, Antonio Zampolli. (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, pp. 1934-1940

Carpuat, Marine and Mona Diab. (2010) Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, CA. Pp. 242-245.

Chafe, Wallace1968. Idiomaticity as an Anomaly in the Chomskyan Paradigm. *Foundations of Language* 4.109-127.

da Silva, Edson Marchetti and Renato Rocha Souza. (2012) Information retrieval system using Multiwords Expressions (MWE) as descriptors. JISTEM - *Journal of Information Systems and Technology Management*. Vol.9 no.2 São Paulo May/Aug. 2012.

Deksne, Daiga, Raivis Skadiņš, and Inguna Skadiņa. a. 2008. Dictionary of Multiword Expressions for Translation into Highly Infliected Languages. In the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco.

Dubremetz, Marie and Joakim Nivre. (2014) Extraction of Nominal Multiword Expressions in French. In proceedings of the 10thWorkshop on Multiword Expressions (MWE 2014), the 14th Conference of the European Chapter of the Association for Computational Linguistics. 26-27 April 2014. Gothenburg, Sweden

Eryiğit, Gülşen, Tugay ¨Ilbay Ozan and Arkan Can. (2011) Multiword Expressions in Statistical Dependency Parsing. Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages.

Fillmore, C.J., P. Kay, M. O'Connor. (1988) Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64, 3, 501–538.

Gadalla, Hassan, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, Cynthia McLemore. (1997) CALLHOME Egyptian Arabic Transcripts. LDC catalog number LDC97T19, ISBN 1-58563-115-9.

Ghoneim, Mahmoud and Mona Diab. (2013) Multiword Expressions in the context of Statistical Machine Translation. In the Proceedings of IJCNLP 2013, October, Nagoya, Japan.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms. *Memory & Cognition*, 8, 449–456.

Grégoire, Nicole. (2010) DuELME: a Dutch electronic lexicon of multiword expressions. In *Language Resources and Evaluation*, 44(1-2):23-39 (2010)

Gross, Maurice, 1986. Lexicon-Grammar. The Representation of Compound Words. In COLING-1986 Proceedings, Bonn, pp. 1-6.

Habash, Nizar, Mona Diab, Owen Rambow (2012). CODA: A Conventional Orthography for Dialectal Arabic. Proceedings of LREC, Istanbul Turkey, May 2012.

Hawwari, Abdelati, Kfir Bar, and Mona Diab (2012). Building an Arabic Multiword Expressions Repository. Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, Montreal, Canada, June 2012.

Jackendoff, R. (1973). The base rules for prepositional phrases. In A *Festschrift for Morris Halle*, pp. 345–356. New York, USA: Rinehart and Winston.

Korkontzelos, Ioannis, and Suresh Manandhar. (2010) Can Recognising Multiword Expressions Improve Shallow Parsing? In proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 636–644, Los Angeles, California, June 2010.

Mel'čuk, I. (1998) Collocations and Lexical Functions. In A.P. Cowie (ed.): *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, 23-53.

Mel'čuk, Igor (2004) Verbes supports sans peine. *Lingvisticæ Investigationes* 27: 2, 203-217.

Nivre, Joakim and Jens Nilsson. 2004. Multiword Units in Syntactic Parsing. InWorkshop on Methodologies and Evaluation of Multiword Units in Real-World Applications, the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 39-46. Lisbon, Portugal.

Odijk, Jan. (2013) Identification and Lexical Representation of Multiword Expressions. In P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing

Palmer, Martha, Dan Gildea, Paul Kingsbury. (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31:1., pp. 71-105.

Ramisch, Carlos, Aline Villavicencio, Christian Boitet, "mwetoolkit: a Framework for Multiword Expression Identification", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta, May, 2010.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP. Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING2002: 1–15.

SanJuan, Eric and Fidelia Ibekwe-SanJuan. 2006. Text mining without document context. In *Information Processing and Management*. Volume 42, Issue 6, pp. 1532-1552.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. (2014) Comprehensive Annotation of Multiword Expressions in a Social Web Corpus. In Proceedings of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, May 2014.

Shudo, Kosho, Akira Kurahone, and Toshifumi Tanabe. (2011) A Comprehensive Dictionary of Multiword Expressions. Proceedings of HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon. Volume 1, pp. 161-170

Walters, Keith. Diglossia, linguistic variation, and language change in Arabic. 1996. In Eid, Mushira, *Perspectives on Arabic Linguistics VIII*. John Benjamins. 1996

Weller, Marion, Ulrich Heid. (2010) Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. In proceedings of the seventh international conference on Language Resources and Evaluation (LREC), Val-letta, Malta.

Zaninello, Andrea, Malvina Nissim. (2010) Creation of lexical resources for a characterisation of multiword expressions in Italian. In proceedings of the seventh international conference on Language Resources and Evaluation (LREC), Valletta, Malta