A Corpus-based Arabic Valency Dictionary: The Case of Fighting Verbs

Esra' M. Abdelzaher *1, Khaled Elghamry**2 , Abeer A. El-Attar***3

*Teaching Assistant, English Language Department, Faculty of Al-Alson
1esraa.abdelzaher@alsun.asu.edu.eg

**Associate Professor, English Language Department, Faculty of Alson
2 elghamryk@gmail.com

***Lecturer, English Language Department, Faculty of Alson
Ain Shams University, Cairo, Egypt
3 elattar@alsun.asu.edu.eg

Abstract: ***Empirical pedagogical dictionaries aim at defining words in their context and presenting corpus-based evidence for each word. They are meant to teach language learners how to use a word correctly. Valency, which describes the arguments of a verb syntactically and semantically, is of unique importance to pedagogical dictionaries. Unfortunately, Arabic lacks corpus-based valency resources. Thus, this paper proposes a monolingual corpus-based valency dictionary, for Arabic learners, covering fighting verbs. The dictionary explores the valency of fighting verbs in Sketch Engine's uploaded Arabic TenTen corpus. The dictionary compiling method depends on both automatic word sketch function to identify the lexico-syntactic patterns of verbs and on three-layer manual annotation of corpus-driven examples to consolidate the results. Each verb entry, in the dictionary, displays (a) number; (b) phrase type; (c) semantic role; (d) grammatical function of its arguments and (e) definition of its different senses. At least, three annotated examples are provided for each verb sense to illustrate its usage authentically. The dictionary, integrating semantic and syntactic information, facilitates effective learning of new Arabic vocabulary.***

Keywords***: Valency dictionaries; Arabic resources; Corpus linguistics; Arabic verbs***

1 INTRODUCTION

A dictionary either theoretically dictates its users what to say and how to say it or describes what is said by a group of people and how it is said [1]. The traditional authoritative approach has been widely adopted in pedagogical dictionaries which aim at teaching learners how to use a linguistic unit correctly. However, the advent of quantitative corpus analysis has strongly undermined the domination of prescriptive dictionaries. Corpus analysis reveals the gap between dictionary language and authentic language, which is actually used by language speakers. Admitting the essentiality of a corpus to dictionary compilation, since the second half of the past century, contemporary dictionaries have been prioritizing corpus data [1].

Teaching Arabic as a second language can be very challenging. Understanding and memorizing Arabic vocabulary are problematic for learners because of inadequate resources [2]. Arabic dictionaries need to overcome certain common problems. For instance, they contain words that are no longer used, keep outdated senses of a word and use inadequate clarifying examples [3]. Using corpus data fixes many of these problems. This paper employs Sketch Engine's TenTen Arabic corpus in compiling a valency dictionary of fighting verbs. The dictionary addresses intermediate Arabic learners. It provides semantic and syntactic information about Arabic verbs and cites corpus-extracted sentences instantiating the proposed lexico-syntactic data.

The remainder of this paper is organized as follows. Section 2 sets the theoretical background of linguistic valency in lexicography. Section 3 reviews the previous application of valency to dictionary construction, highlighting valency exploration of Arabic verbs in particular. Section 4 details the methodology steps of creating the proposed lexicon. Section 5 displays the results and exemplifies the final lexicon entries. It also discusses the challenges and applicability of the employed methodology. Finally, the last section presents the concluding remarks and the future work recommendations.

2 THEORETICAL BACKGROUND

A. Valency Definition

Valency, in chemistry, denotes a chemical combination capacity of one element to be in bond with a number of atoms. Similarly, valency, in linguistics, refers to the ability of a verb, or any word, to have one or more role participant. For instance, *die* is univalent. It typically combines with one participant, which is the one who dies. *Kill*, however, is divalent as it typically involves a killer and a dead person. Some verbs are trivalent (e.g., give). The verb *give* requires the existence of three participants, giver, taker and a given object [4].

So, the basic and broad sense of valency is related to the number and type of verb arguments. Verb valency in general is related to the semantic roles played by the verb subject and object, their phrase type, optional and obligatory valents, among others. However, valency is differently used within the scope of Case Grammar and Function Generative Grammar. Each grammar approach views valency from a certain perspective. Thus, different terms are used to denote very similar concepts. Thematic roles, semantic roles, participant roles and frame elements almost refer to the same thing, which is the semantic relation between the subject and the object, for instance, and the verb in a sentence. Studying the valency of verbs enables further sub-classification of verbs, according to lexico-syntactic features. Thus, valency pays little attention to general verb characteristics, such as tense and mood. The grammatical label *subject* can correspond to a set of semantic roles, including agent, experiencer and caustor. Accordingly, assigning the semantic role to the grammatical function of a word significantly contributes to its meaning. Assigning the semantic role to the verb participants gives insights into more semantic details revealing the nature of the verb itself [5].

This paper studies valency within the Case Grammar framework. Valency is defined as 'the particular kinds of constituents, in terms of semantic roles, grammatical functions, and phrase types, with which a word combines in a grammatical sentence' [6]. The target words, in this paper, are Arabic verbs of fighting.

B. Valency Triangle

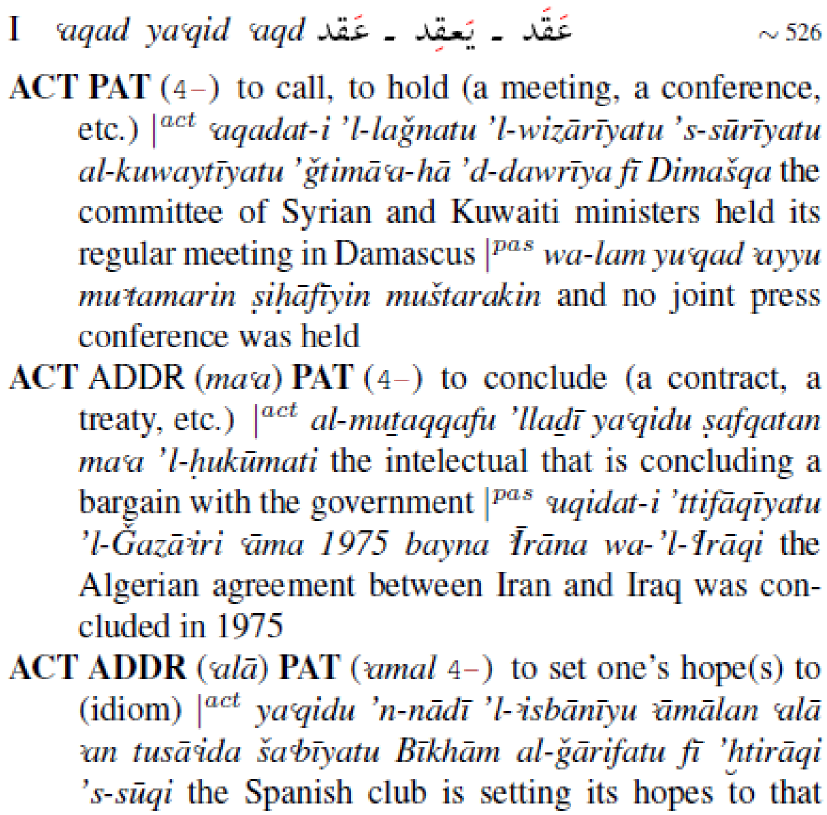
Valency is not only patterns of association between a target word and other words or phrases. It is a semantic or syntactic 'requirement' which should be met; otherwise, the sentence is 'unacceptable' [7]. Valency is dependent on three basic terms: valency group, pattern and description. *Valency group,* which embraces a frame element and its phrase type and grammatical function, refers to the three entities as general categories. Their realization, however, forms the *valency pattern* of a sentence or a construction. If a frame element is grammatically labeled as subject, semantically tagged as an agent and syntactically instantiated by a noun phrase, a valency pattern of Subj-Agent-NP is identified. *Valency description* is the sum of all valency patterns within which a target word occurs [8]. The presented dictionary applies this triangle in the annotation process in order to construct the final verb entry.

C. Tri-layered Annotation

Valency is divided into syntactic and semantic valency. Syntactic valency is concerned with the syntactic elements occurring with the target word in a construction while semantic valency is related to the conceptual situation associated with the target word and the participants involved in such situation [6]. Thus, a three-layer annotation is needed to conclude the valency of a verb. At the syntactic level, verb participants are tagged as subject, direct object, indirect object, genitive determiner, among others. Then, the phrase type of each instantiated participant is identified and added as an annotation layer. At the semantic level, each verb is placed within a certain frame or situation. Polysemous verbs are placed within their respective frames. Then, the frame elements of each verb are identified. Semantic and syntactic annotations integrate each other. To elaborate, 'circumstances forced the doctor to treat her enemies', if syntactically tagged, provides no significant facts about the target verb 'treat'. It only labels 'circumstances' as 'subject' and does not refer to 'doctor' as a doer of any action. The semantic analysis, however, labels doctor as the external argument of 'treat', despite playing no active/subject role grammatically. It is marked as 'external' because it is not mentioned within the headword's phrase; otherwise it is labeled 'dependent' [9].

3 RELATED WORKS

Very few attempts are made to build valency Arabic dictionaries. An Arabic valency lexicon explores the most frequent verbs. It uses the annotated corpus of Prague Arabic Dependency Treebank and applies valency within the Function Generative Grammar approach. The outcome entry includes root, lexeme, derivational class, equivalent verbal noun, lexical unit, valency frame, frequency information, among others. For instance, عقد/hold-conclude-set hope has three valency frames corresponding to three lexical units. First, the sense of holding a meeting has the agent/patient participants. Second, the sense of concluding a contract corresponds to the addressee/patient participants. Patient, in this case, co-occurs with the preposition مع/with. Finally, the sense of setting hope corresponds to the addressee/patient participants. However, patient in the third sense collocates with the preposition على/on [10]. However, the lexicon uses English as a representation language and relies heavily on transliteration. Only the verb is written in Arabic, followed by a transliteration. Then, the different valency patterns are written in English and a suggested English translation is provided. Even the Arabic clarifying examples are displayed in transliteration and translation. This affects the usability of the lexicon by Arabic learners. Figure 1 visualizes a partial entry of the lexicon.



I *ʿaqad yaʿqid ʿaqd* عَقَد ـ يَعقِد ـ عَقْد ~ 526

ACT PAT (4–) to call, to hold (a meeting, a conference, etc.) $|^{act}$ *ʿaqadat-i ’l-lağnatu ’l-wizārīyatu ’s-sūrīyatu al-kuwaytīyatu ’ǧtimāʿa-hā ’d-dawrīya fī Dimašqa* the committee of Syrian and Kuwaiti ministers held its regular meeting in Damascus $|^{pas}$ *wa-lam yuʿqad ʾayyu muʾtamarin ṣiḥāfīyin muštarakin* and no joint press conference was held

ACT ADDR (*maʿa*) **PAT** (4–) to conclude (a contract, a treaty, etc.) $|^{act}$ *al-muṯaqqafu ’llaḏī yaʿqidu ṣafqatan maʿa ’l-ḥukūmati* the intelectual that is concluding a bargain with the government $|^{pas}$ *ʿuqidat-i ’ttifāqīyatu ’l-Ǧazāʾiri ʿāma 1975 bayna ʾĪrāna wa-’l-ʿIrāqi* the Algerian agreement between Iran and Iraq was concluded in 1975

ACT ADDR (*ʿalā*) **PAT** (*ʾamal* 4–) to set one’s hope(s) to (idiom) $|^{act}$ *yaʿqidu ’n-nādī ’l-ʾisbānīyu ʾāmālan ʿalā ʾan tusāʿida šaʿbīyatu Bīkhām al-ǧārifatu fī ’ḫtirāqi ’s-sūqi* the Spanish club is setting its hopes to that

Figure1. Verb entry in a valency Arabic lexicon

This lexicon [10] and the proposed dictionary in this study, despite addressing the same linguistic phenomenon, are fundamentally different. First, the lexicon [10] adopts the Function Generative Grammar approach to valency. It views the syntactic patterns and verb collocates as set of optional or obligatory conditions under which a verb can occur. The proposed dictionary, following the Case Grammar valency approach, starts with the lexico-syntactic patterns and ends with the conceptual frame associated with a verb. Second, the lexicon methodologically starts with annotated corpus, which facilitates linguistic pattern identification. The proposed dictionary starts with a classic Arabic resource to gather seed words and drives information from general reference Arabic corpus and modern dictionary. Moreover, the lexicon selects the analyzed sample of verbs quantitatively, while the proposed dictionary resorts to a qualitatively-chosen sample.

Accordingly, the results of the two resources differ. As displayed in figure 1, verb entry in the lexicon [10], despite dividing word meaning into senses, does not explicitly identify any of the associated frames. Lacking clear identification of verb frames prevents the lexicon from grouping cognitively related word senses together. The proposed dictionary, giving the heaviest weight to semantic and cognitive information, accompanies the list of verb senses with their frames and frame participants. Moreover, the lexicon extensively uses English translation and Arabic transliteration. It also abbreviates English semantic roles in its annotated examples. The proposed dictionary uses Arabic only to display the results and avoids abbreviated forms as displayed in table 2.

Still, the lexicon [10] covers a broader verb scope than the proposed dictionary which targets a set of words. It covers the most frequent Arabic verbs and their derivational forms. Data resources and verb entries, however, are richer in the proposed dictionary. Moreover, the structure of the lexicon requires an intermediate English knowledge from a user. This is beneficial for translators and learners having Arabic and English background information. The proposed dictionary lacks the English translation provided by the lexicon [10]. However, it can be used by any Arabic learner.

Linguists may resort to completely manual construction of valency dictionaries, at the early stages of the work, in case of lacking large syntactically annotated corpus of the target language. This case is applicable on the Romanian verb valency lexicon. The lexicon is the outcome of 3 years’ work of 5 linguists. It covers 3000 verbs extracted from print dictionaries. It also uses other resources to identify the different senses of the verb. Still, it verifies the dictionary information in a newspaper corpus [11]. Manual effort seems to be mandatory in compiling or developing valency dictionaries. To illustrate, the Czech valency dictionary of verbs is a project lasted for more than one year. All the annotated sentences in the lexicon database are manually tagged [12]. Similarly, developing a Czech verb valency dictionary involves manual filtration of automated results and manual annotation of corpus-driven examples [13]. The analyzed words in valency dictionaries are syntactically chosen according to their grammatical functions, i.e. nouns, verbs or adjectives. However, most valency dictionaries focus primarily on verbs [10-15]. Verbs can be selected quantitatively, because of their high ranking in the frequency list of general reference corpora, or qualitatively, depending on their semantic field

4 METHODOLOGY

A. Corpus Description

TenTen Arabic Corpus is one of Sketch Engine web-based corpora for major languages in the world. TenTen refers to the corpus size. TenTen corpora are cleaned and duplicated at the paragraph level [16]. TenTen Arabic corpus is created in 2012 to be a source of contemporary written Arabic language. The corpus is 5.8 million words covering a variety of domains across different Arab countries [17]. Now, the corpus is available online on Sketch Engine website. It is tagged by Stanford Arabic Parser in 2015.

B. List of Fighting Verbs

Target words in the dictionary are, at the grammatical level, verbs. The dictionary focuses, in this preliminary stage, only on verbs for a number of reasons. First, valency originally targets verbs but its scope has extended to include nouns and adjectives. Second, verbs represent the original and richest grammatical category for valency analysis. Third, identifying the pattern within which a verb occurs syntactically and semantically reveals the deep and surface meaning of a sentence. Fourth, nouns and adjectives in Arabic represent derivational forms of the verb. Thus, identifying the valency of fighting verbs is essential to identify the valency of fighting nouns and adjectives.

The analyzed verbs are qualitatively selected. Pre-determining the semantic field of the target verbs, especially at the early stages of construction, facilitates the process of different senses disambiguation and helps in filtering the patterns within which a verb occurs. So, fighting verbs are chosen as the topic of the valency dictionary.

A combination of classic and contemporary Arabic resources is used to form the list of verbs. Seed words are extracted from a classical Arabic thesaurus [18]. The keyword حارب/fight is searched and all its synonyms are extracted as candidate seed words. Then, the verbs are bootstrapped in the TenTen corpus to return more modern synonymous verbs. The results are manually filtered to extract verbs of fighting, exclude irrelevant verbs and group non-verb words, if related to the fighting domain, under their corresponding verbs. This step should enable future research to be extended to fighting nouns and adjectives.

C. Creating Verb Database

Before writing the verb entry, a database containing all verb information is constructed. First, a verb is searched in a modern Arabic dictionary [19]. It is employed for its user-friendly electronic interface, valuable information concerning word combinations and contemporary illustrative examples. However, to be entered in a database, dictionary verb information is organized into senses, collocates and examples. Second, the same verb is explored in TenTen corpus to verify or discard the provided dictionary information. Word sketch function is employed to gather the lexico-syntactic collocates of each verb. The collocates are classified according to their grammatical categories to separate the different senses of the verb, if any. This step is based on the conclusion that different verb senses occur in different syntactic patterns [20]. The collocates and their grammatical labels are checked against the dictionary senses in the created database to create an initial one-to-one or one-to-many correspondence between a sense and a pattern. They are also re-checked in the corpus to extract the full linguistic context in a 10-window-size concordance. The corpus-driven examples are annotated and added to the database. These final two steps either authenticate the suggested sense-pattern correspondence or nullify it. The database contains (a) dictionary defined senses of a verb; (b) dictionary collocates; (c) dictionary illustrative sentences; (d) corpus-driven lexico-syntactic patterns; (e) initially concluded patterns based on the grammatical category of the collocates and (f) corpus-driven annotated examples.

D. Annotation Process

Annotation is a process of adding ‘interpretive linguistic information’ to a text [21]. In the proposed dictionary, two types of explaining details are needed. First, semantic annotation of corpus examples is performed. A ten-window-size concordance is automatically extracted via Sketch Engine. The concordance is processed, considering the dictionary senses added to the database, to select examples for annotation. For each verb sense, a frame and frame elements are assigned. That is to say, the conceptual situation, in which each verb sense is used, is added as a semantic information layer. Moreover, the typical semantic roles played by the typical participants in the defined situation are also tagged. Second, a syntactically-annotated layer is added to the interpretive information of a target verb. At the syntactic level, the grammatical function and the phrase type of each verb sense are identified. Thus, the different patterns of a verb are syntactically and semantically figured out.

E. *Writing Entries*

After creating the database, only valency-related information is obtained to be displayed in a verb entry. Lexico-syntactic collocates, unless empirically consolidated as part of a pattern, are excluded. Senses that have no authentic existence are also discarded. New corpus-driven senses are added. The final entry is displayed in simple intermediate-level Arabic. It includes (a) verb senses; (b) corresponding frames; (c) involving frame elements; (d) semantic-syntactic patterns and (e) annotated examples.

Figure 2 outlines the methodology of constructing the proposed dictionary. It starts with targeting a keyword in order to collect fighting seed words. Then, it displays the process of searching for the seed words in both Arabic dictionary and corpus. It highlights the three types of information entered in the database: dictionary information, manual annotated examples and corpus-driven data. Finally, it refers to the last step: writing verb entry.

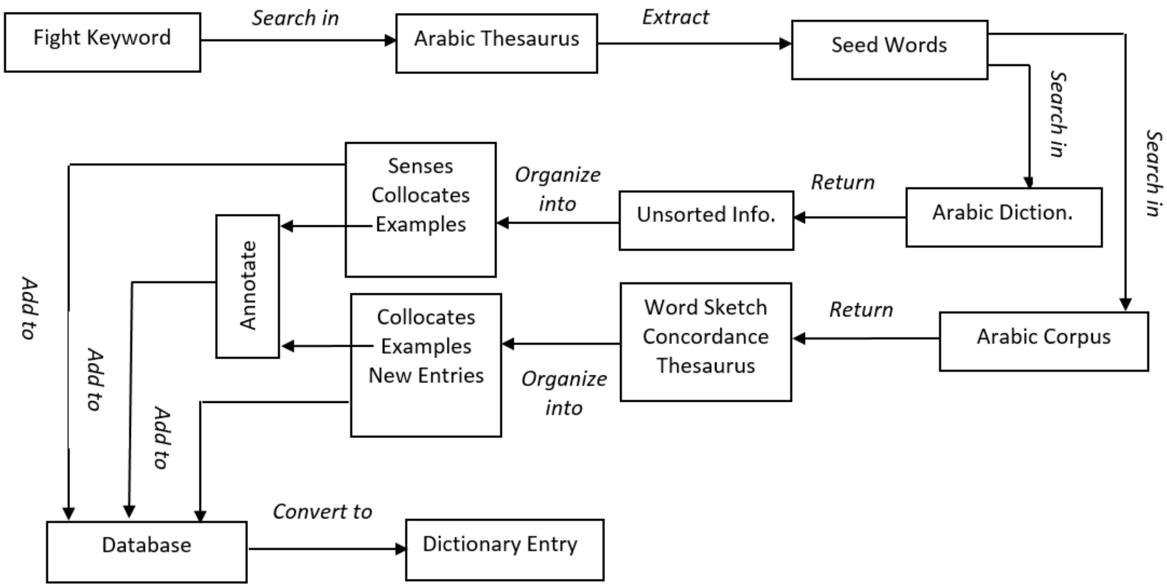
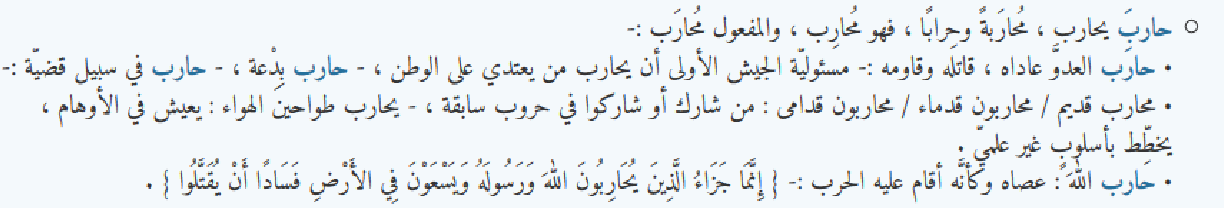


Figure2. Methodology of creating the proposed dictionary

5 Results and Discussion

A. *Dictionary information in the constructed database*

Abdul Hamid's "Dictionary of Contemporary Arabic Language" displays a variety of information about a verb in a very simplified way. However, entering such information in a database requires semantic and syntactic reorganization and filtration. Figure 3 displays the dictionary entry of حارب/fought verb.



○ حارَبَ يحارِبِ ، مُحارَبةً وحِرابًا ، فهو مُحارِب ، والمفعول مُحارَب :-

- حارب العدوّ عاداه ، قاتله وقاومه :- مسئوليّة الجيش الأولى أن يحارب من يعتدي على الوطن ، - حارب بِدْعة ، - حارب في سبيل قضيّة :-
- محارب قديم / محاربون قدماء / محاربون قدامى : من شارك أو شاركوا في حروب سابقة ، - يحارب طواحينَ الهواء : يعيش في الأوهام ، يخطّط بأسلوب غير علمي .
- حارَب اللهَ : عصاه وكأنّه أقام عليه الحرب :- { إِنَّمَا جَزَاءُ الَّذِينَ يُحَارِبُونَ اللهَ وَرَسُولَهُ وَيَسْعَوْنَ فِي الأَرْضِ فَسَادًا أَنْ يُقَتَّلُوا } .

Figure3. حارب/fought entry in a modern Arabic dictionary

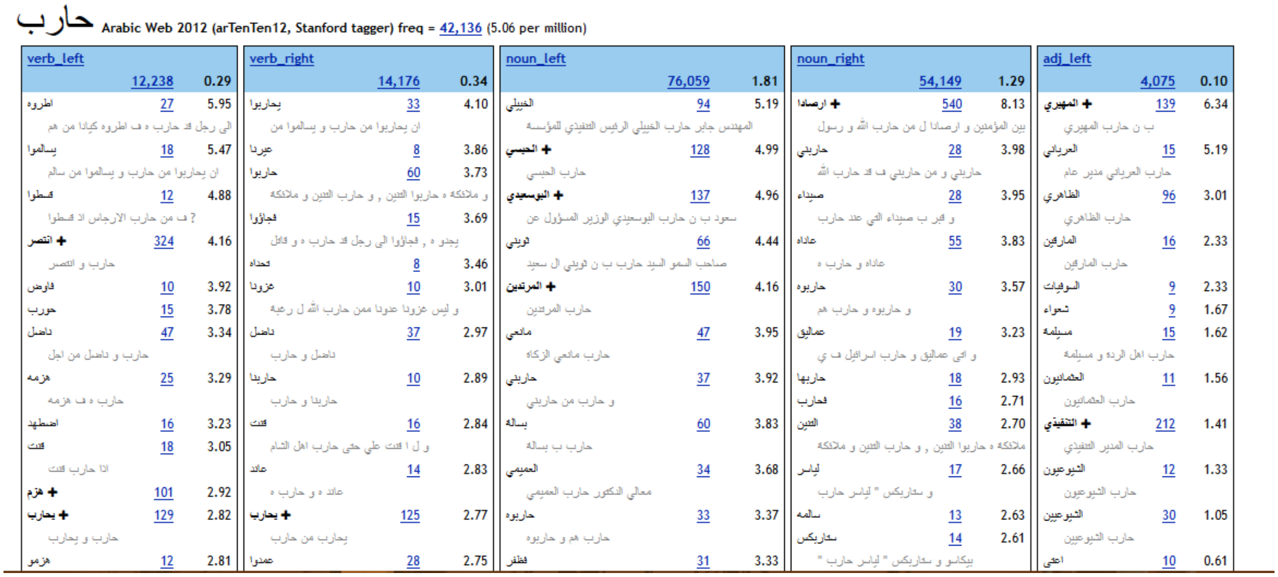
Eliciting valency-related information is done through extracting different senses, collocates, figurative expressions and illustrative examples. Classifying dictionary information into the previously-mentioned categories undergoes some interpretive additions. The semantic roles of the collocates, which are candidate frame elements, are manually added to the database. Syntactic and grammatical labels are also outlined. A distinction between literal and figurative meanings, if any, is made. حارب/fought in the database is represented as follows:

Table 1
DICTIONARY INFORMATION OF حارب/FOUGHT IN THE CONSTRUCTED DATABASE

أمثلة	الأدوار الدلالية	نمط لغوي	تلازم لفظي	معنى	فعل
مسئوليّة الجيش الأولى أن يحارب من يعتدي على الوطن	المنفذ	اسم> فاعل	الجيش	عادى، قاتل، قاوم	حارب
	الضحية	اسم> مفعول به	العدو		
	الغاية	شبه جملة> جار ومجرور	في سبيل		
إِنَّمَا جَزَاءُ الَّذِينَ يُحَارِبُونَ اللهَ وَرَسُولَهُ وَيَسْعَوْنَ فِي الأَرْضِ فَسَادًا أَنْ يُقَتَّلُوا		اسم> مضاف إليه	الله (تلازم إلزامي)	عصى (معنى مجازي)	
		اسم> مفعول به			
		اسم مضاف> مفعول به	طواحين الهواء (تلازم إلزامي)	يعيش في الأوهام. (معنى مجازي)	

B. Corpus-driven Information

Word sketch function does not work in Arabic as effectively as in English. However, it effectively contributes to validating dictionary information, revealing more synonyms, determining frames and concluding frame elements. حارب/fought word sketch empirically authenticates the three senses of the verb displayed in the dictionary. It also suggests ناضل, جاهد, غزا and كافح as synonyms of حارب. Word sketch presents war, struggle, hatred, resistance, among others, as potential frames within which حارب can be used. Figure 4 visualizes a partial word sketch of حارب/fought.



حارب Arabic Web 2012 (arTenTen12, Stanford tagger) freq = 42,136 (5.06 per million)

verb_left	12,238	0.29
اطروه	27	5.95
الي رجل قد حارب ه ف اطروه كيادا من هم		
يسالموا	18	5.47
ان يحاربوا من حارب و يسالموا من سالم		
قسطوا	12	4.88
ف من حارب الارجاس لا قسطوا ?		
انتصر +	324	4.16
حارب و انتصر		
فاوض	10	3.92
حورب	15	3.78
ناضل	47	3.34
حارب و ناضل من اجل		
هزمه	25	3.29
حارب ه ف هزمه		
اضطهد	16	3.23
قتت	18	3.05
اذا حارب قتت		
هزم +	101	2.92
يحارب +	129	2.82
حارب و يحارب		
هزمو	12	2.81

verb_right	14,176	0.34
يحاربوا	33	4.10
ان يحاربوا من حارب و يسالموا من		
عيرنا	8	3.86
حاربوا	60	3.73
و ملائكة ه حاربوا التنين , و حارب التنين و ملائكة		
فجاؤوا	15	3.69
يجدو ه , فجاؤوا الي رجل قد حارب ه و قاتل		
تحداه	8	3.46
غزونا	10	3.01
و ليس غزونا عدونا ممن حارب الله ل رغبة		
ناضل	37	2.97
ناضل و حارب		
حاربنا	10	2.89
حاربنا و حارب		
قتت	16	2.84
و ل ا قتت علي حتي حارب اهل الشام		
عاند	14	2.83
عاند ه و حارب ه		
يحارب +	125	2.77
يحارب من حارب		
عمدوا	28	2.75

noun_left	76,059	1.81
الخيلي	94	5.19
المهندس جابر حارب الخيلي الرئيس التنفيذي للمؤسسة		
الحبسي +	128	4.99
حارب الحبسي		
اليوسعيدي +	137	4.96
سعود ب ن حارب اليوسعيدي الوزير المسؤول عن		
ثويني	66	4.44
صاحب السمو السيد حارب ب ن ثويني ال سعيد		
المرتدين +	150	4.16
حارب المرتدين		
مانعي	47	3.95
حارب مانعي الزكاة		
حاربني	37	3.92
و حارب من حاربني		
بسالة	60	3.83
حارب ب بسالة		
العميمي	34	3.68
معالي الدكتور حارب العميمي		
حاربوه	33	3.37
حارب هم و حاربوه		
قطقر	31	3.33

noun_right	54,149	1.29
ارصادا +	540	8.13
بين المؤمنين و ارصادا ل من حارب الله و رسول		
حاربني	28	3.98
حاربني و من حاربني ف قد حارب الله		
صيداء	28	3.95
و قبر ب صيداء التي عند حارب		
عاداه	55	3.83
عاداه و حارب ه		
حاربوه	30	3.57
و حاربوه و حارب هم		
عماليق	19	3.23
و اتي عماليق و حارب اسرائيل ف ي		
حاربها	18	2.93
فحارب	16	2.71
التنين	38	2.70
ملائكة ه حاربوا التنين , و حارب التنين و ملائكة		
ليسار	17	2.66
و ستاريكس " ليسار حارب		
سالمه	13	2.63
ستاريكس	14	2.61
بيكاسو و ستاريكس " ليسار حارب "		

adj_left	4,075	0.10
المهيري +	139	6.34
ب ن حارب المهيري		
العرياني	15	5.19
حارب العرياني مدير عام		
الظاهري	96	3.01
حارب الظاهري		
المارقين	16	2.33
حارب المارقين		
السوفيات	9	2.33
شعواء	9	1.67
مسيلمة	15	1.62
حارب اهل الردة و مسيلمة		
العثمانيون	11	1.56
حارب العثمانيون		
التنفيذي +	212	1.41
حارب المدير التنفيذي		
الشيوعيون	12	1.33
حارب الشيوعيون		
الشيوعيين	30	1.05
حارب الشيوعيين		
اعمي	10	0.61

Figure 4: Partial word sketch of حارب/fought in TenTen corpus

C. Integrating dictionary and corpus-driven information

The suggested information displayed in the word sketch is checked in a 10-word concordance of the verb to (in)validate the observations driven from the word sketch. The extracted sentences are annotated to illustrate the frame, frame elements, sense, syntactic and semantic patterns of a verb. At least one annotated example is added to each verb sense. The final entry, in the dictionary, is demonstrated in the table 2.

Table 2
SAMPLE OF VERB ENTRIES IN THE PROPOSED DICIONARY

أمثلة	**النمط اللغوي**	**تكافؤ الفعل**	**عناصر الإطار**	**الإطار**	**معنى**	**فعل**
أبو بكر_**مبتدأ.منفذ**_*حارب* المرتدين_**مفعول به.ضحية**_	اسم>مبتدأ>منفذ اسم>مفعول به>ضحية	ثنائي	منفذ ضحية	حرب	عادى وقاتل	حارب
الجزائر هذه ليست تلك التي *حارب* أباؤنا وأجدادنا_**فاعل.منفذ**_ من أجلها_**غاية**_ فرنسا_**مفعول به.ضحية**_	اسم>فاعل>منفذ اسم>مفعول به>ضحية جار ومجرور>غاية	ثلاثي	منفذ ضحية غاية	مقاومة	قاوم	
قال أنه_**اسم أن.منفذ**_ *يغزو* العراق_**مفعول به_موضوع**	ضمير متصل>اسم أن> منفذ اسم> مفعول به> موضوع	ثنائي	منفذ موضوع	هجوم	هاجم وسار إلى قتال	غزا
اكتشف أن الشيب_**اسم أن.منفذ** قد *غزا* الرأس_**مفعول به.ضحية**_	اسم>اسم أن> منفذ اسم>مفعول به> ضحية	ثنائي	منفذ ضحية	انتشار	تكاثر وانتشر	
النظام الجديد_**مبتدأ.منفذ**_*يغزو* الأسواق_**مفعول به.موضوع**_	اسم>اسم أن> منفذ اسم>مفعول به> موضوع	ثنائي	منفذ موضوع			

The proposed dictionary integrates existing Arabic dictionary knowledge with authentic Arabic corpus. Using both in a dictionary compilation is evidently effective. First, the used dictionary contains semantic and syntactic details which are essential to the proposed lexicon. However, preprocessing such details is requisite before adding them to a database. The used dictionary depends on synonymous definition of the target verbs. This technique is valuable as it enriches the proposed dictionary with new entries. However, it does not provide detailed definition of each sense. The constructed database cites every example illustrated in the dictionary. Processing the data before adding them to the valency database involves manual addition of the grammatical function, phrase type and semantic role of the extracted collocate.

Second, the recruited corpus is the source of valency information of the target verbs. Exploring verbs in Arabic TenTen corpus reveals promising results. Word sketch and concordance lines play significant role in validating dictionary information and reaching the ultimate goal of determining verb valency. They also enrich the entries with contemporary synonyms. The corpus captures each dictionary sense, even the figurative ones. Starting from the existing dictionary entries enables filtering word sketch results via suggesting possible patterns for senses. It also informally supervises selecting sentences for annotation depending on the candidate dictionary senses.

Checking the final verb entry in the proposed lexicon against classic and contemporary Arabic dictionaries [18], [20], [23], [24] and [25] is significantly necessary to experiment the lexicon. First, at the morphological level, Arabic dictionaries usually include inflectional and derivational forms of a word in the entry. However, subject-sorted dictionaries, which are similar to the proposed dictionary in the thematic sorting of words, lack this morphological feature. They marginalize morphological information for the sake of semantic details. In a further developmental step of the dictionary, the morphological information added to the verb database should be incorporated in the verb entry. Second, at the semantic level, classic dictionaries give illustrative detailed definitions of words and countless examples. However, they group all word senses and sentence examples together. This makes adding such information to a database very exhaustive. Contemporary dictionaries, despite providing very economic definitions, exert some effort to separate different senses of a word and provide at least one example for each sense. This dictionary is similar to contemporary dictionaries in concisely defining word but it adds more explanatory corpus-based examples for each word sense to compensate for the brief definitions. Third, the proposed dictionary, at the grammatical level, breaks the norm of contemporary dictionaries in which grammatical information are relegated. It identifies the grammatical functions of verb participants to highlight some grammar-in-use issues for Arabic learners. Classic dictionaries extensively tackle grammatical issues but they stipulate advanced grammatical knowledge to follow up. The proposed dictionary is grammatically between classic and modern dictionaries. Fourth, unlike other dictionaries, the dictionary addresses the cognitive level of language. It considers sense frame identification to be essential to any definition. Finally, the adopted methodology returns retrievable results starting from the list of seed words extracted from the classic thesaurus [18] and ending with the annotated examples cited from TenTen corpus.

6 Conclusions and Future work

To conclude, this paper constructs a database, of fighting verbs, based on the Contemporary Arabic Dictionary and Arabic TenTen corpus. The database is used to construct a valency dictionary of fighting verbs, which is the main objective of the paper. Building a valency dictionary of Arabic fighting verbs is motivated by the lack of Arabic valency resources in general, and corpus-based valency dictionaries in particular. The proposed dictionary, covering verbs of fighting, provides corpus-based semantic and syntactic valency information about verbs. It aims to teach Arabic learners how to use verbs correctly. Unlike traditional Arabic dictionaries, it provides annotated examples to elaborate (a) different senses; (b) frames; (c) frame elements of a verb and (d) grammatical function; (e) phrase type of each frame element. Accordingly, a detailed valency description of each verb is identified. Semantic field valency analysis is very fruitful at the early stages of building dictionaries.

The constructed valency database of fighting verbs contains additional information, such as lexico-syntactic collocates and derivational verb forms, which can be used in future research. Future work may extend to nouns, adjectives and other word classes belonging to the fighting semantic field. Being verb-driven, fighting nouns and adjectives are supposed to carry similar semantic features to that of verbs. The same corpus-based methodology would be applied to conclude their syntactic patterns, which are essentially different from verbs', and compare their semantic features to verbs. Finally, adding valency information to monolingual dictionaries would enhance the performance of translators and machine translation systems.

REFERENCES

[1] B. T. Atkins, and Michael Rundell. *The Oxford guide to practical lexicography*. Oxford University Press, 2008.

[2] M. Yaakub "Teaching Arabic as a second language: An evaluation of key word method effectiveness." *Sains Humanika* Vol. 46, No. 1, 2007.

[3] J. Halpern. "Compilation Techniques for Pedagogically Effective Bilingual Learners' Dictionaries." *International Journal of Lexicography* Vol. 29, No. 3 , pp. 323-338, 2016.

[4] "What Is Valency?" URL: <<http://www-01.sil.org/linguistics/glossaryoflinguisticterms/WhatIsValency.htm>>. Retrieved on 09 Feb, 2017.

[5] O. Smrz, P. Zemánek, J. Krácmar, and V. Bielický. "Information Structure with the Prague Arabic Dependency Treebank.", 2006.

[6] C. J. Fillmore, and M. Petruck. "Frame Net glossary." *International Journal of Lexicography* Vol. 16, No. 3, pp.359-361, 2003.

[7] C. J. Fillmore, C. Johnson, and M Petruck. "Background to Frame Net." *International journal of lexicography* Vol. 16, no. 3, pp. 235-250, 2003.

[8] J. Ruppenhofer, M. Ellsworth, M. R. Petruck, C. Johnson, and J. Scheffczyk. "FrameNet II: Extended theory and practice.", 2006.

[9] S. Atkins, C. J. Fillmore, and C. Johnson. "Lexicographic relevance: Selecting information from corpus evidence." *International Journal of Lexicography*. Vol. 16, No. 3, pp. 251-280, 2003.

[10] V. Bielický, and O. Smrz. "Building the Valency Lexicon of Arabic Verbs." In *LREC,* 2008.

[11] Barbu, Ana-Maria. "First Steps in Building a Verb Valency Lexicon for Romanian." In *International Conference on Text, Speech and Dialogue*, pp. 29-36. Springer, Berlin Heidelberg, 2008.

[12] Lopatková, M., Řezníčková, V. and Žabokrtský, Z. Valency Lexicon for Czech: From Verbs to Nouns. In *International Conference on Text, Speech and Dialogue* (pp. 147-150). Springer Berlin Heidelberg, 2002.

[13] Skoumalová, H., Straňáková-Lopatková, M. and Žabokrtský, Z. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In *International Conference on Text, Speech and Dialogue* (pp. 142-149). Springer Berlin Heidelberg, 2001.

[14] Kotsyba, N. Using Polish Wordnet for Predicting Semantic Roles for the Valency Dictionary of Polish Verbs. In *International Conference on Natural Language Processing* (pp. 202-207). Springer International Publishing, 2014.

[15] Kettnerová, V., Lopatková, M. and Hrstková, K. Semantic roles in Valency lexicon of Czech verbs: Verbs of communication and exchange. In *Advances in Natural Language Processing* (pp. 217-221). Springer Berlin Heidelberg, 2008.

[16] M. Jakubíček, A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. The TenTen corpus family. In *7th International Corpus Linguistics Conference CL* (pp. 125-127), 2013.

[17] A. Tressy, Y. Belinkov, N. Habash, A. Kilgarriff, and V. Suchomel. "arTenTen: Arabic corpus and word sketches." *Journal of King Saud University-Computer and Information Sciences* 26, no. 4 pp. 357-371, 2014.

[18] A. I. Hamadhani: book of words (Investigated by: Badrawi Zahran). Dar Almaaref Publishing House. 3rd Ed. Cairo. Egypt.

[20] A. M. Abdul Hamid "Dictionary of contemporary Arabic language". World of Alkottob, 1st Ed., Cairo, Egypt, 2008.

[21] S. Atkins, M. Rundell, and H. Sato. "The contribution of Framenet to practical lexicography." *International Journal of Lexicography* Vol. 16, No. 3 pp. 333-357, 2003.

[22] G. Leech. "Introducing corpus annotation." (1997).

[23] A. A. Rummani. Near synonymous words (Investigated by: Fath-Allah Al-Masry). Dar Alwafaa Publishing House. Mansoura, Egypt, 1987.

[24]M. Manzur. Lisan Alarab. Dar ṣādir, bdūn tarīkḫ. Almujlalad. Beirut, Lebanon, 1994.

[25] The Academy of the Arabic Language. Al-Wassit. 4th Ed. Cairo, Egypt, 2004.

BIOGRAPHY

Esra M. Abdelzaher is a Teaching Assistant, Faculty of Alson, Ain Shams University, Egypt. Her research interests include corpus and cognitive linguistics.

Dr. Khaled Elghamry is an Associate Professor of (Computational) Linguistics, Faculty of Alson, Ain Shams University, Egypt. Dr. Elghamry obtained a Ph.D. in Computational Linguistics, 2004, Indiana University, USA. He was a visiting scholar, Department of Linguistics, University of Florida 2007-2010. Dr. Elghamry is the Co-Founder of the Midwest Computational Linguistics Forum, Indiana University, and also the Co-Founder of the Arabic Digital Content Statistical Database, and the Arabic Digital Content Foundation Report. Dr. Elghamry presented and published research in international conferences and journals on different technical issues in the automated processing of the Arabic language and content, Weband text-mining, sentiment analysis, and tracking of online public opinion.

Dr. Abeer El-Attar is a Lecturer of Linguistics, Faculty of Alson, Ain Shams University, Egypt. Dr.El-Attar obtained a Ph.D. in Linguistics, 2000.

TRANSLATED ABSTRACT

دراسة ذخائرية لبناء قاموس تكافؤ معاني لأفعال القتال في اللغة العربية

1 اسراء عبد الظاهر، 2 خالد الغمرى، 3 عبير العطار
قسم اللغة الأنجليزية - كلية الألسن – جامعة عين شمس
[1]esraa.abdelzaher@alsun.asu.edu.eg
[2]elghamryk@gmail.com
[3]elattar@alsun.asu.edu.eg

ملخص:

تهدف الدراسة إلى استخدام المنهج التجريبي في بناء معجم خاص بأفعال القتال في اللغة العربية. تعتمد الدراسة على ذخيرة لغوية مكونة من 5.8 مليون كلمة عربية فصحى معاصرة، وتشتمل الذخيرة على نصوص عربية مجمعة من الإنترنت من كافة المجالات والبلاد العربية. يقدم المعجم تعريفًا، لكل فعل، مبني على سياقات استخدامه الواردة في الذخيرة اللغوية، كما يركز -بشكل أساسي- على نظرية تكافؤ المعاني. تعتمد النظرية على حصر السياقات النحوية والدلالية لكل فعل، وصولًا إلى مجموعة الأنماط التي يمكن أن يرد بها فعل ما. تعد نظرية التكافؤ من أهم النظريات المستخدمة في إطار تعلم اللغة؛ فهي تساعد المتعلمين على ربط الكلمات بالسياقات اللغوية التي تستخدم بها؛ مما ييسر اكتساب الكلمات الجديدة، وتوظيفها توظيفًا لغويًا سليمًا.

الكلمات الدلالية: علم الذخائر اللغوية، تكافؤ المعاني، القواميس العربية، الأفعال العربية