A Lexical Resource of Hebrew Verb-Noun Multi-Word Expressions

Chaya Liebeskind, Yaakov HaCohen-Kerner
Department of Computer Science, Jerusalem College of Technology, Lev Academic Center
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel
liebchaya@gmail.com, kerner@jct.ac.il

Abstract

A *verb-noun Multi-Word Expression* (MWE) is a combination of a verb and a noun with or without other words, in which the combination has a meaning different from the meaning of the words considered separately. In this paper, we present a new lexical resource of Hebrew *Verb-Noun MWEs* (VN-MWEs). The VN-MWEs of this resource were manually collected and annotated from five different web resources. In addition, we analyze the lexical properties of Hebrew VN-MWEs by classifying them to three types: morphological, syntactic, and semantic. These two contributions are essential for designing algorithms for automatic VN-MWEs extraction. The analysis suggests some interesting features of VN-MWEs for exploration. The lexical resource enables to sample a set of positive examples for Hebrew VN-MWEs. This set of examples can either be used for training supervised algorithms or as seeds in unsupervised bootstrapping algorithms. Thus, this resource is a first step towards automatic identification of Hebrew VN-MWEs, which is important for natural language understanding, generation and translation systems.

Keywords: Hebrew, Lexical Resource, Multi-Word Expression (MWE), Verb-Noun Multi-Word Expression

1. Introduction

A *Multi-Word Expression* (MWE) is defined as any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts (Sag et al., 2002). Examples of MWEs are phrasal verbs (*calm down*, *get around*), compounds (*bus stop*, *washing machine*), and idioms (*break a leg*, *raining cats and dogs*). The numerous MWEs pose a great challenge to the creation of *Natural Language Processing* (NLP) systems (Biber et al., 1999). They play an important role in NLP applications, such as semantic parsing and machine translation, which should not only identify them, but also deal with them when they are encountered (Fazly and Stevenson, 2007).

In this research, we focus on a particular class of Hebrew MWEs that are formed from the combination of a verb with a noun. Each Hebrew *Verb-Noun MWE* (VN-MWE) functions as a single semantic unit and usually has an idiomatic meaning not predictable from the meanings of the individual parts.

Al-Haj (2009) presented a systematic linguistic characterization of MWEs in Hebrew, and provided in a full picture of the diverse properties that Hebrew MWEs exhibit. However, Al-Haj and Wintner (2010) limited their investigation to noun compounds. Although verb-noun MWEs acquisition has been widely investigated in English, as far as we know, our research is a first attempt to focus specifically on this MWE type in Hebrew.

The goal of our research is to manually construct a high quality publishable lexical resource of Hebrew VN-MWEs, which is a useful tool for supporting automatic extraction of additional VN-MWEs. Such a tool is important due to the fact that despite our efforts to make the lexical resource as comprehensive as possible, there are VN-MWEs that are not covered by our resource. Moreover, writers and speakers often create new VN-MWEs that are not included in any lexical resource. For example, "I had *Googled out* a relevant website" means I managed to find a relevant web site by using the Google search engine.

An additional contribution of the ongoing research presented in this paper is an analysis of the properties of Hebrew VN-MWEs. This analysis is a first necessary step for designing algorithms for automatic VN-MWEs extraction. The analysis suggests some interesting features of VN-MWEs for exploration. Then, a set of examples can be sampled from the lexical resource. This set can either be used for training supervised algorithms or as seeds in unsupervised bootstrapping algorithms.

In Section 2, we aim to provide the necessary background needed for the subsequent sections and for placing this work within the line of other works on MWEs extraction in Hebrew. Section 3 elaborates on the linguistic properties related to Hebrew VN-MWEs. In Section 4, firstly, we describe our lexical resource. Then, we analyze the distribution of the linguistic properties over the resource's entries. In Section 5, we suggest directions for future research.

2. Background

In general, approaches to automatic identification of MWEs can be divided into three categories: 1. Statistical approaches based on frequency and co-occurrence affinity (Dias et al., 1999; Deane, 2005; Pecina and Schlesinger, 2006). 2. Linguistic approaches using parsers, lexicons and language filters (Al-Haj and Wintner, 2010; Bejcek et al., 2013; Green et al., 2013; Al-Haj et al., 2014), and 3. Hybrid approaches combining different methods (Baldwin, 2005; Zhang et al., 2006; Fazly, 2007; Boulaknadel et al., 2008; Ramisch et al., 2010; Farahmand and Nivre, 2015; Sangati and van Cranenburgh, 2015).

Considerable research has been done on automatic identification of MWEs in English (Venkatapathy and Joshi, 2004; Schneider et al., 2014), German (Breidt, 1996) and other European languages (Dandapat et al., 2006; Todirascu and Navlea, 2015) but not much research have been carried at this level for Hebrew.

Al-Haj (2009) presented a systematic linguistic characterization of MWEs in Hebrew. However, Al-Haj and Wintner (2010) limited their experiment to noun compounds. They classified MWE candidates to compounds and non-compounds by a supervised learning approach, and showed that relying on linguistic information dramatically improves the accuracy of compound extraction.

Tsvetkov and Wintner (2012) proposed an algorithm for identifying MWEs in bilingual corpora, using automatic word alignment as their main source of information. However, automatic construction of parallel corpora is a time consuming task.

Later, Tsvetkov and Wintner (2014) proposed a novel architecture for identifying MWEs of various types and syntactic categories by Bayesian network models in monolingual corpora. Their approach addressed MWEs of various types by zooming in on the general idiosyncratic properties of MWEs rather than on specific properties of each subclass thereof. Addressing multiple types of MWEs has its limitations: The task is less well-defined, one cannot rely on specific properties of a particular construction, and the type of the MWE is not extracted along with the candidate expression.

Recently, Fadida et al. (2014) presented a verb-complement dictionary of Modern Hebrew, automatically extracted from text corpora. We plan to investigate the utilization of this dictionary for our verb-particle identification task along with other available Hebrew lexical resources. Although verb-particle MWEs acquisition has been widely investigated in English, as far as we know, our suggested research is a first attempt to focus specifically on this MWE type in Hebrew.

3. Lexical resource of Hebrew Verb-Noun Multi-Word Expressions

Our lexical resource currently includes 505 entries of VN-MWEs. It was manually constructed from 5 web resources: the Hebrew language dictionary[1], wiktionary's idioms category[2], wiktionary's idioms from the Talmud and the Mishnah category, wiktionary's idioms from the Bible category, and the free Hebrew dictionary in the Web[3]. We limited our resource to VN-MWEs of length two and three words (247 bigrams and 258 trigrams). We evaluated the inter-annotator agreement over 150 candidate terms that were randomly sampled from our web resources. Two annotators judged their suitability for our lexical resource. We observed a Kappa (Cohen, 1960) value of 0.86, which is considered as almost perfect (Landis and Koch, 1977). Table 1 summarizes the most frequent part-of-speech (POS) patterns in our lexical resource. The distribution of the patterns has a long tail of patterns that appear less than 10 times. The POS were automatically assigned by a tagger (Adler, 2007; Adler et al., 2008). About 20% of the VN-MWEs were tagged incorrectly, assessing the necessity of our manual annotation. Examples for frequent types of tagging errors are presented in Table 2. We note that prepositions and definite articles are often prefixes in Hebrew. The lexical resource is publicly available for download in uft8 format[4]. Each entry includes the VN-MWEs, its lemmatized form and POS pattern, as were assigned by the tagger.

4. Linguistic properties of Hebrew Verb-Noun Multi-Word Expressions

Following Al-Haj (2009), we classify the linguistic properties of Hebrew VN-MWEs along three dimensions: morphological, syntactic, and semantic. In Table 3 (see the last page), we shortly cite the description of the properties in each of these categories, as defined by Al-Haj et al. (2014), and provide examples of Hebrew VN-MWEs from our lexical resource for each case. These properties will be used to distinguish between VN-MWEs and verb-noun combinations, which are not MWEs.

Table 4 presents the distribution of the lexical properties of Hebrew VN-MWEs over a sample of 100 VN-MWEs from our lexical resource. For each property, the percentage is calculated separately. The most characteristic properties of VN-MWEs are the semantic properties of compositionality and lexical fixedness, 92% of the VN-MWEs have a high degree of idiomaticity. While the VN-MWEs' syntax tends to be fixed, 82% of the VN-MWEs do not allow any changes in the constituent order and 87% are non-compositional, their morphology mostly allows partial inflection. These findings suggest that focusing on the exploration of the VN-MWEs semantic properties is a promising research direction.

The distribution of lexical properties (%)		
Morphological	Frozen form	17
	Partial inflection	71
	Hapax legomena	5
Syntactic	Compositionality	13
	Constituent order	18
Semantic	Semantic compositionality	92
	Lexical fixedness	94
	Translation equivalents	23

Table 4: The distribution of the lexical properties of Hebrew VN-MWEs

5. Future work

We plan to use the lexical resource to construct a dataset of examples for supervised algorithms. Then, we plan to investigate algorithms for classifying collocations that include verbs as VN-MWEs or non-VN-MWEs. These algorithms would combine three types of features, inspired by the three lexical properties that we have discussed in the

[1] https://www.safa-ivrit.org/milon_map.php

[2] https://he.wiktionary.org/wiki

[3] http://milog.co.il/

[4] http://liebeskind-chaya.blogspot.co.il/p/downloads.html

[5] To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexico-graphic order, are abgdhwzxTiklmns`pcqršt.

[6] The first word is a tagging error of the first type

previous section. For instance, the semantic compositionality property can be modeled by co-occurrence and distributional similarity measures. Moreover, we plan to sample Hebrew VN-MWEs from our resource and use them as seeds in unsupervised bootstrapping algorithms for classification of collocations.

6. References

Adler, M., Goldberg, Y., Gabay, D., and Elhadad, M. (2008). Unsupervised lexicon-based resolution of unknown words for full morphological analysis. In *Proceedings of ACL-08: HLT*, pages 728–736, Columbus, Ohio, June. Association for Computational Linguistics.

Adler, M. (2007). *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. Ph.D. thesis.

Al-Haj, H. and Wintner, S. (2010). Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Al-Haj, H., Itai, A., and Wintner, S. (2014). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Al-Haj, H. (2009). *Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition*. Ph.D. thesis, University of Haifa.

Baldwin, T. (2005). Deep lexical acquisition of verb-particle constructions. *Comput. Speech Lang.*, 19(4):398–414, October.

Bejcek, E., Stranák, P., and Pecina, P. (2013). Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 106–115.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman grammar of spoken and written English*, volume 2. MIT Press.

Boulaknadel, S., Daille, B., and Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In *LREC*. European Language Resources Association.

Breidt, E. (1996). Extraction of vn-collocations from text corpora: A feasibility study for german. *arXiv preprint cmp-lg/9603006*.

Dandapat, S., Mitra, P., and Sarkar, S. (2006). Statistical investigation of bengali noun-verb (nv) collocations as multi-word-expressions. *the Proceedings of MSPIL*, pages 230–233.

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 605–613. Association for Computational Linguistics.

Dias, G., Guilloré, S., and Lopes, J. P. (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *Traitement Automatique des Langues Naturelles, Institut dEtudes Scientifiques, Cargèse, France*, pages 333–339.

Fadida, H., Itai, A., and Wintner, S. (2014). A hebrew verbcomplement dictionary. *Language Resources and Evaluation*, 48(2):249–278.

Farahmand, M. and Nivre, J. (2015). Modeling the statistical idiosyncrasy of multiword expressions. In *Proceedings of NAACL-HLT*, pages 34–38.

Fazly, A. and Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16. Association for Computational Linguistics.

Fazly, A. (2007). *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, University of Toronto.

Green, S., de Marneffe, M.-C., and Manning, C. D. (2013). Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227, March.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 651–658, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ramisch, C., de Medeiros Caseli, H., Villavicencio, A., Machado, A., and Finatto, M. J. (2010). A hybrid approach for multiword expression identification. In Thiago Alexandre Salgueiro Pardo, et al., editors, *PROPOR*, volume 6001 of *Lecture Notes in Computer Science*, pages 65–74. Springer.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK, UK. Springer-Verlag.

Sangati, F. and van Cranenburgh, A. (2015). Multiword expression identification with recurring tree fragments and association measures. In *Proceedings of NAACL-HLT*, pages 10–18.

Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Todirascu, A. and Navlea, M. (2015). Aligning verb+noun collocation to improve a french-romanian statistical mt system. *Multi-word Units in Machine Translation and Translation Technology*.

Tsvetkov, Y. and Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Tsvetkov, Y. and Wintner, S. (2014). Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Venkatapathy, S. and Joshi, A. K. (2004). Recognition of multi-word expressions: A study of verb-noun (vn) collocations. In *Proceedings of the International Conference on Natural Language Processing*, volume 2004.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

#	Part-of-speech pattern	Relative freq.	Example
1	verb+noun	16.04%	*rah awr* (lit. ”saw light”) ”to be published”
2	verb+preposition noun	9.9%	*šins mtniw* (lit. ”to gird his waists”) ”to buckle down”
3	verb+preposition+definiteArticle noun	8.91%	*akl at hkwb`* (lit. ”ate the hat”) ”publicly admitted his mistake”
4	verb+preposition+noun	8.71%	*qpch `liw zqnh* (lit. ”old-age jumped on him”) ”premature aging”
5	verb+noun+preposition noun	2.77%	*šlx id bnpšw* (lit. ”sent a hand in his soul”) ”to commit suicide”
6	verb+preposition definiteArticle noun	2.57%	*ird mharc* (lit. ”went down the land”) ”to emigrate from Israel”
7	verb+preposition noun+noun	2.18%	*išb baps m`šh* (lit. ”sat with zero act”) ”to be idle”

Table 1: The most frequent POS patterns in our lexical resource (the plus sign separates between the MWE’s constituents)

#	Tagging error	Relative freq.	Examples	Ambiguous word in the example
1	verb as noun	41.3%	*Tb` xwtmw*[5] (noun+noun) (lit. ”coined his seal”) ”to leave one’s impression” *npx at npšw* (noun+preposition+noun) (lit. ”blew his soul”) ”to breathe one’s last”	*Tb`* ”to coin” v.s ”nature” *npx* ”to blow” v.s ”capacity”
2	noun as properName	12.5%	*qra drwr* (verb+properName) (lit. ”call freedom”) ”to free” *nša brkh* (verb+properName) (lit. ”carried a greeting”) ”congratulated”	emphdrwr (flowery) ”freedom” *brkh* ”greeting”
3	verb as properName	11.5%	*dn lkp zkwt* (properName+preposition noun+noun) (lit. ”judged to the cape of a good deed”) ”to judge someone favorably” *asp bzrw`wtiw* (properName+preposition noun) (lit. ”collected in his arms”) ”to hug”	*dn* ”to judge” *asp* ”to collect”
4	noun as verb	7.7%	*akl xcc* (verb+verb) (lit. ”ate gravel”) ”to have a difficult time” *la lqx lrawt* (lit. ”did not take to the lungs”) (negation+verb+verb) ”smoking without inhaling the smoke to the lungs”	*akl xcc* ”gravel” v.s ”to divide” *lrawt* ”lung” v.s ”to see”
5	verb as adjective	6.7%	*mgdlim pwrxim bawir* (noun+adjective+preposition definiteArticle noun) (lit. ”castles in the sky”) ”imaginary plans” *axztw xmh* (noun+adjective)[6] (lit. ”the sun reached him”) ”he has fever”	*pwrxim* ”to fly” vs. ”flying away” *xmh* ”sun” v.s ”her heat”

Table 2: Frequent types of tagging errors

Category	**Property**	**Description**	**Example**
Morphological	Frozen form	Constituents can appear in one fixed (frozen) form	Citation (canonical) form: *mˋz ica mtwq* (lit. ”from strength came out sweet”) ”all for the best”. Frozen inflected form: the words *gdiim* and *tiišim* (the plural form of *gdi* (lit. ”young goat”) and *tiišim* (lit. ”billy goat”) in *gdiim nˋšw tiišim* (lit. ”young goats became billy goats”) ”grow up”.
	Partial inflection	Constituents undergo a (strict) subset of the full inflections that they would undergo in isolation	the verb *akl* (lit. ate) in the expression *akl bkl ph* (lit. ”(he) ate with all his mouth”) can inflect for number, gender, person, and all tenses. However, the noun *ph* does not inflect for number. So, it does not appear in the form ”they ate with all their mouths”.
	Hapax legomena	Constituents that have no other usage or literal meaning outside the expression they appear in	*akl kwrca* (lit. ”ate a piece of meat” ”to slander”. The second word *kwrca*, by itself, has no literal meaning in modern Hebrew (it is an Aramaic word).
Syntactic	Compositionality	VN-MWEs that contain *open slots*, which can be filled with complements of certain parts of speech	*dxh bqš* (lit. ”reject with straw”) ”to evade a question”. The open slot must be filled by a prepositional phrase, as in *dxh at xbrw bqš* (lit. ”reject his friend with straw”) ”to evade his friends’ question”.
	Constituent order	A number of syntactic structures that, when used compositionally, permit a change in the order of constituents while preserving the meaning of the whole expression	*nddh šntw* (lit. ”his sleep wandered”) ”unable to fall asleep” the order of the words can be replaced, *šntw nddh* is a valid noun-verb expression.
Semantic	Compositionality	The degree to which the meaning of the whole expression results from combining the meanings of its individual words when they occur in isolation	the VN-MWEs *rwah at hnwld* (lit. ”see the born”) ”to anticipate” and *hdˋt swblt* (lit. ”the mind suffers”) ”unacceptable” have a high degree of idiomaticity.
	Lexical fixedness	Replacing any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or a literal expression	An exception is the word *nwtr* ”left” in the verb-noun expression *nwtr ˋl knw* (lit. ”left on its base”) ”to stay unchanged” that can be replaced by the word *ˋmd* ”stand”.
	Translation equivalents	VN-MWEs that translate, as a whole, to a single word in some other language, or literally to a verb-noun expression in the target language (English in our case)	Translate as a whole: *kwbš at icrw* (lit. ”conquers his urge”) ”overcome”. Translate literally: *hpk at ˋwrw* ”to change one’s skin”

Table 3: Examples of the linguistic properties of Hebrew VN-MWEs (The three left columns are taken from Al-Haj et al. (2014))