Towards a classification of Hebrew verbal MWEs

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner

WG1

1 Overview

As part of our ongoing project of developing linguistically motivated, deep grammars of Hebrew and Arabic (Arad Greshler et al., 2015; Herzig Sheinfux et al., 2015, Forthcoming), we are focusing our attention on incorporating verbal MWEs into the lexicon and the grammars. Our goal is to identify a large number of such expressions, classify them according to the syntactic processes that they can undergo, and extend the grammar with appropriate constructions for handling them, as the need arises.

We consider various types of syntactic "transformations" that were mentioned in the literature (e.g., Nunberg et al., 1994): internal modification, anaphora, ellipsis, topicalization/focalization, quantification, word order variations, and pluralization. In addition, we investigate the possibility of substituting words with synonyms. Our main research question is whether MWEs can be classified according to the processes that they can undergo, and more specifically, whether there are correlations between the syntactic flexibility of MWEs and their semantic opacity.

To better address this question, we focus first on Hebrew MWEs which include words or constituents that have no usage outside the particular expression (*fossil words*). As fossil words are typically semantically opaque, MWEs that include them are likely to be opaque as well. Our hypothesis is that such MWEs will tend to be syntactically rigid. A clear advantage of focusing on such expressions is that their unique distributional property makes them especially suitable for corpus searches, with the fossil word serving as the search term, or pivot.

Consequently, the research questions that we focus on with respect to this particular class of MWEs are: (1) Does the opaque semantics of fossil words in MWEs correlate with more conservative usage patterns? (2) Are there sub-classes within the set of syntactic "transformations" which pattern together with respect to these MWEs? To address these questions we use *heTenTen 2014*, a billion-token web-crawled Hebrew corpus (Baroni et al., 2009) that is morphologically analyzed and disambiguated. The corpus is loaded into SketchEngine (Kilgarriff et al., 2004), which allows sophisticated queries to be posed.

2 Preliminary results

We are currently compiling a list of verbal MWEs with fossil words. Examples include MWEs with fossil verbs and nouns (1), MWEs headed by a fossil verb (2), and MWEs with fossil nouns (3).

(1) *hem lo margiʃim ʃe-hem* ***magdiʃim*** *et ha-****seʔa***
they not feel that-they over.fill ACC the-biblical.measure
'They don't feel that they are overdoing it.'

(2) *paʕam nosefet* ***hiqdiħa*** *moʕecet ha-menahalim et ha-tavʃil*
*instance additional burned.*3SF *board the-directors ACC the-cooked.food*
'Once again the board of directors caused damage.'

In what follows we focus on our findings regarding the MWE *ʔavad ʕalav (ha-)kelaħ* '*outdated, obsolete*' (3). It is made up of three components: the verb *ʔavad* '*lost*', a PP headed by the preposition *ʕal* '*on*', and the fossil noun *kelaħ*, whose meaning is opaque (and hence lacks a gloss in the following examples). This noun, which functions as the subject in the expression, can either be definite (3a) or indefinite (3b). The preposition *ʕal* '*on*' can either take a pronominal clitic (3a) or a full NP (3b).

(3) a. *yeʃ ʕaraxim ʃe-ʔavad ʕaley-hem* ***kelaħ***
exist values.PM *that-lost*.3SM *on-them*.PM *KELAH*
'There are values that are outdated.'

b. *ʔavad ha-****kelaħ*** *ʕal kol ha-nimuqim ha-ʔele*
lost.3SM *the-KELAH on all the-excuses the-these*
'All these excuses are obsolete.'

The word *kelaħ* is a bona fide fossil word, as it is not attested in Modern Hebrew outside of this expression and its meaning is unclear. Consequently, the expression cannot have a figurative interpretation. An additional peculiarity is that the verb *ʔavad* '*lost*', a relatively frequent verb, appears with a PP complement headed by *ʕal* '*on*' only in the context of this expression.

Nevertheless, in spite of its opaque semantics and unusual complement selection, this expression is not rare. Its frequency is 1.5 per million in *heTenTen 2014* (cf. *kick the bucket* which occurs 0.1 per million in COCA (Davies, 2008-). Interestingly, unlike its biblical forefather, in Modern Hebrew it is mostly used with a definite *kelaħ* (75% of occurrences).

The special characteristics of this expression, as well as its semantically unary denotation ('become outdated') led us to expect it to exhibit conservative usage patterns. Corpus searches, however, reveal that in spite of these circumstances the expression does exhibit some variability. One example of such variability is word order alternations. As shown in (3) above, the expression can appear in two verb-initial word orders, VOS and VSO, both found in Modern Hebrew, and associated with non-topical subjects (Melnik, 2006). An additional construction is illustrated in (4), where a focalized PP appears clause-initially.

(4) *ʕal ha-hagdara ha-zot ʔavad ha-****kelaħ***
on the-definition the-this lost.3SM *the-KELAH*
'This definition is obsolete.'

Interestingly, one word order in which this expression rarely occurs is SVO, the unmarked word order in Modern Hebrew. The reason for this lacuna may be that an opaque fossil word is incompatible with this position, which is associated with a prototypical topical subject; the expression cannot be *about* the KELAH. Nevertheless, the following example was attested.

(5) *mosad geriyatri* [$_{\#1}$ *ʃe-ha-****kelaħ*** [$_{\#2}$ *ʃe-ʔavad ʕalav*] *kvar heħlid mizman*]
institution geriatric that-the-KELAH that-lost on.it already rusted long.ago
'a geriatric institution that has long been outdated.'

In (5) the KELAH is the subject of the relative clause, tagged #1 and is, in turn, modified by an inner relative clause, tagged #2, which contains the MWE. The predicate in clause #1, *heħlid* '*rusted*', reveals the speaker's conceptualization of the referent of KELAH – an object which rusts with age.

We focused the discussion on one particular MWE, nevertheless at this stage in our investigation we already see more syntactic flexibility with regards to MWEs with fossil words than we had expected. Our immediate goal is to expand our dataset and to obtain a clearer picture of the interaction of semantic opacity with syntactic flexibility.

References

Arad Greshler, Tali, Livnat Herzig Sheinfux, Nurit Melnik & Shuly Wintner. 2015. Development of maximally reusable grammars: Parallel development of Hebrew and Arabic grammars. In Stefan Müller (ed.), *Proceedings of the 22nd international conference on Head-driven Phrase Structure Grammar*, 27–40. Stanford, CA: CSLI Publications.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources And Evaluation* 43(3). 209–226. `http://www.springerlink.com/index/10.1007/s10579-009-9081-4`.

Davies, Mark. 2008-. The corpus of contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Herzig Sheinfux, Livnat, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. 2015. Hebrew verbal multi-word expressions. In Stefan Müller (ed.), *Proceedings of the 22nd international conference on Head-driven Phrase Structure Grammar*, 122–135. Stanford, CA: CSLI Publications.

Herzig Sheinfux, Livnat, Nurit Melnik & Shuly Wintner. Forthcoming. Representing argument structure. Journal of Linguistics.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In *Proceedings of EURALEX*, 105–116.

Melnik, Nurit. 2006. A constructional approach to verb-initial constructions in Modern Hebrew. *Cognitive Linguistics* 17(2). 153–198.

Nunberg, Geoffrey, Ivan A. Sag & Tom Wasow. 1994. Idioms. *Language* 70. 491–538.