Hebrew Verbal Multi-Word Expressions

Livnat Herzig Sheinfux[1], Tali Arad Greshler[1], Nurit Melnik[2] and Shuly Wintner[1]

[1]Department of Computer Science, University of Haifa
[2]Department of Literature, Language and the Arts, The Open University of Israel

1 Introduction

Multi-word expressions (MWEs) in Modern Hebrew (MH), as in other languages, are not simple to characterize, since they vary in the degree of idiosyncrasy with respect to their semantic, syntactic, and morphological behavior. In this study we focus on verbal MWEs: we consider different types of this class of MWEs, and propose an analysis in the framework of HPSG (Pollard and Sag, 1994). Moreover, we incorporate this analysis in HeGram (Herzig Sheinfux et al., 2015), a deep linguistic processing grammar of Modern Hebrew.

Our motivation is twofold. First, the need to incorporate MWEs into the grammar is unquestionable, especially in light of estimates claiming that MWEs account for approximately half of the entries in the lexicon (Sag et al., 2002). Second, we view MWEs as a challenging test case for the innovative architecture implemented in HeGram.

2 Verbal MWEs in Hebrew

The syntactic idiosyncrasies involved in verbal MWEs make their incorporation into linguistically-motivated grammars challenging. We identify the following patterns of MWEs in MH:

- MWEs can be headed by verbs which lexically select for a particular NP complement (1) or for a PP headed by a particular preposition and complemented by a particular NP (2):

  (1) *dan higdil roš*
  Dan made.grow head
  'Dan took.on responsibility.'

  (2) *dan yarad me-haʿec*
  Dan went.down from-the.tree
  'Dan conceded.'

- Some MWEs are headed by verbs which select for possessive NPs, either as complements of the verb (3) or as complements in the PP complement of the verb (4), and impose agreement between the possessor and one of the verb's dependents:

  (3) *ha-anašim$_i$ ṭamnu yad-am$_i$*
  *the-people buried.*3P *hand-their*
  ba-calaχat
  in.the-plate
  'The people refrained from acting.'

  (4) *dan$_i$ yaca mi-kelav$_i$*
  Dan came.out from-tools.his
  'Dan lost his temper.'

- MWEs can include "empty slots", filled by non-idiomatic and unrestricted complements (e.g., *Dana* in (5))

(5) *dan hoci et dana$_i$ me-ha-kelim / mi-keleiha$_i$*
Dan took.out ACC Dana from-the-tools / from-tools.her
'Dan made Dana lose her temper.'

- MWEw can sometimes allow an idiomatic complement to be internally modified:

(6) *ha-cibur nafal ba-pax̲ (ha-pirsumi)*
the-public fell in.the-bin the-advertising
'The public was tricked (by advertisement).'

We account for these and similar MWEs in an HPSG grammar of MH. In what follows we will outline our solution to the following challenges: (i) Representing both literal and idiomatic instances of expressions; (ii) Capturing selectional restrictions; (iii) Enabling internal modification; and (iv) Accounting for non-local selection.

3 The incorporation of MWEs into the grammar

3.1 HeGram

Our proposed analysis is cast in the context of HeGram (Herzig Sheinfux et al., 2015), a deep linguistic processing grammar of Modern Hebrew, which is implemented in the LKB (Copestake, 2002) and ACE systems. HeGram was developed in parallel with AraGram (see Arad Greshler et al., 2015), a grammar of Modern Standard Arabic, and it is based on a corpus survey of the 50 most frequent verbs in Hebrew (100 instances each).

The architecture of the grammar embodies significant changes to the way argument structure is standardly viewed in HPSG. Specifically, it distinguishes between semantic selection and syntactic selection, and provides a way of stating constraints regarding each level separately. Moreover, one lexical entry accounts for multiple subcategorization frames, including argument optionality and the realization of arguments with different syntactic phrase types (e.g., *want food* vs. *want to eat*). This architecture is similar in spirit to work done on Polish by Przepiórkowski et al. (2014).

The VALENCE features in HeGram are distributed across ten categories. Each valence category is characterized in terms of its semantic role, as well as the types of syntactic phrases which can realize it. Consequently, the semantic relations denoted by predicates consist of coherent argument roles, which are consistent across all predicates in the language. See a fuller discussion in Herzig Sheinfux et al. (2015).

3.2 Representing both literal and idiomatic instances of expressions

One phenomenon which receives a comprehensive account in HeGram is multiple subcategorization, whereby predicates appear in a number of different subcategorization frames. In a way, MWEs constitute an extreme case of this type of variability. Verbs which head VP MWEs can occur in 'standard' VP constructions, as well as in idiomatic ones. The degree of overlap between the behavior of the verb in its standard guise and in its idiomatic role is mostly verb-specific. Nevertheless, regardless of the degree, our lexical inheritance hierarchy enables us to distinguish between shared properties and those which differ in the two instantiations.

For example, the verb *hoci* ('*take.out*') in (5) semantically selects two complements: *Theme* and *Source*. Syntactically, the two complements are realized as NP and PP, respectively. These properties are shared by both the literal and the idiomatic verb, and consequently the two senses inherit from a common supertype which specifies this information.

The two senses diverge in a number of ways. As expected, the idiomatic sense is more restrictive in terms of its selectional restrictions. The *Source* argument can only be a PP headed by a specific selective preposition *mi* (‘*from*’), which in turn is complemented by an NP headed by the idiomatic plural definite noun *ha-kelim* (‘*the tools*’). Moreover, the *Source* NP can optionally appear with a possessor suffix, provided that it is co-indexed with the *Theme* argument of the verb. Any divergence from these restrictions eliminates the idiomatic reading. The literal *hoci* (‘*take.out*’) selects a *Source* PP that is headed by the standard preposition *mi* (‘*from*’). With the literal sense, however, the *Source* argument is optional (e.g., *hoci et ha-sefer* (‘*took out the book*’)). The optional realization of arguments is captured in HeGram by the specification of realization frames, which indicate which semantic arguments are obligatory and which are optional.

It is worth mentioning that our analysis does not distinguish decomposable from non-decomposable idioms, as we only have a relatively superficial semantic representation of MWEs. All the idiomatic components of an MWE have separate idiomatic entries in the lexicon, which include their idiomatic meaning.

3.3 Selectional restrictions

MWEs require various degrees of flexibility in lexical selection. For example, the MWE in (2) requires the NP complement to be headed by the singular noun *ʿec* (‘*tree*’), regardless of whether the subject is singular or plural. In (3), however, plural subjects can either bury their singular *hand* or plural *hands* in the (singular) *plate*. This, of course, is expression-specific and needs to be specified in the lexicon.

Moreover, although a word in an idiomatic expression is identical in form to its literal counterpart, their semantic content is distinct. In order to distinguish between literal and idiomatic words, and to control their distribution, semantic relations are divided into *l(iteral)-rel*s and *i(diomatic)-rel*s (Copestake, 1994; Sag et al., 2002; Kay and Sag, 2012, among others). Idiomatic lexemes like *hoci* (‘*take.out*’) (in (5)) have semantic *i-rel*s. They select for a *Source* PP with a specific idiomatic *_from_tools_ip_rel* relation. This selective preposition *me* (or *mi*), in turn, selects for an NP with an idiomatic *_i-tools-temper_n_rel* relation. This notwithstanding, the *Theme* argument of the idiomatic *hoci* (‘*take.out*’) is an “open slot” and can be filled by any NP complement, provided that it is not idiomatic (i.e., has an *l-rel*).

3.4 Non-local selection

Verbal MWEs exhibit two types of non-local selection phenomena. First, verbs which lexically select the complements of their PP complements do so indirectly by selecting specific prepositions, which select specific idiomatic complements. Consequently, the lexicon includes multiple lexical entries for these selective “idiomatic” prepositions, each pertaining to a different MWE in which they appear.

Indirect lexical selection such as the one described above, where a verb selects for a preposition which selects for a noun, forms a type of a chain, where heads of phrases select heads of other phrases. This mechanism is supported by the TOPREL feature, an independently motivated feature in HeGram, which identifies the main semantic relation denoted by a lexeme. Idiomatic selectors target this feature (see Fig. (1)), which percolates from head daughter to the “mother” phrase.[1] Moreover, the TOPREL feature provides a way of allowing for internal modification, e.g., (6); the TOPREL of a modified phrase is identical to the main relation of the head.

Admittedly, using a selectional chain to ensure that idiomatic verbs that select specific PPs only combine with the correct complements introduces some redundancy to the lexicon. However, this solution does solve the non-local selection problem, and the TOPREL feature we use to implement it is independently motivated in HeGram.[2]

[1] Kay and Sag (2012) suggest a similar feature, LEXICAL-ID (LID).

[2] Although there is no independent evidence for the existence of an idiomatic form of prepositions, usage patterns diverge:

A second type of non-local selection involves co-indexed possession. For example, in (5) the possessor of the NP complement in the *Source* PP *mi-keleiha* (‘*from-her.tools*’) must be co-indexed with the *Theme* NP *Dana*. Consequently, this index must be “visible” at the PP level. The feature which projects the lower possessor to this higher level is the XARG feature (Kay and Sag, 2012; Bond et al., 2013).

Different idiomatic MWEs have different patterns of co-indexed possession, so the exact structure-sharing pattern is lexically specified per verb type.[3] For example, while in (4) the possessor of the NP complement inside the PP must be co-indexed with the subject (i.e., the XARG of the *Source* is structure-shared with the XARG of the verb), in (5) the possessor of the NP complement inside the PP must be co-indexed with the NP complement (i.e., the XARG of the *Source* is structure-shared with the INDEX of the *Theme*).

3.5 Illustration

The example sentence in (5) poses most of the challenges described above. It is an “empty slot” MWE, with an idiomatic PP complement with a possessed NP whose possessor is obligatorily co-indexed with the literal NP complement filling the “slot”. Figure 1 illustrates the essential components of our analysis of this MWE.

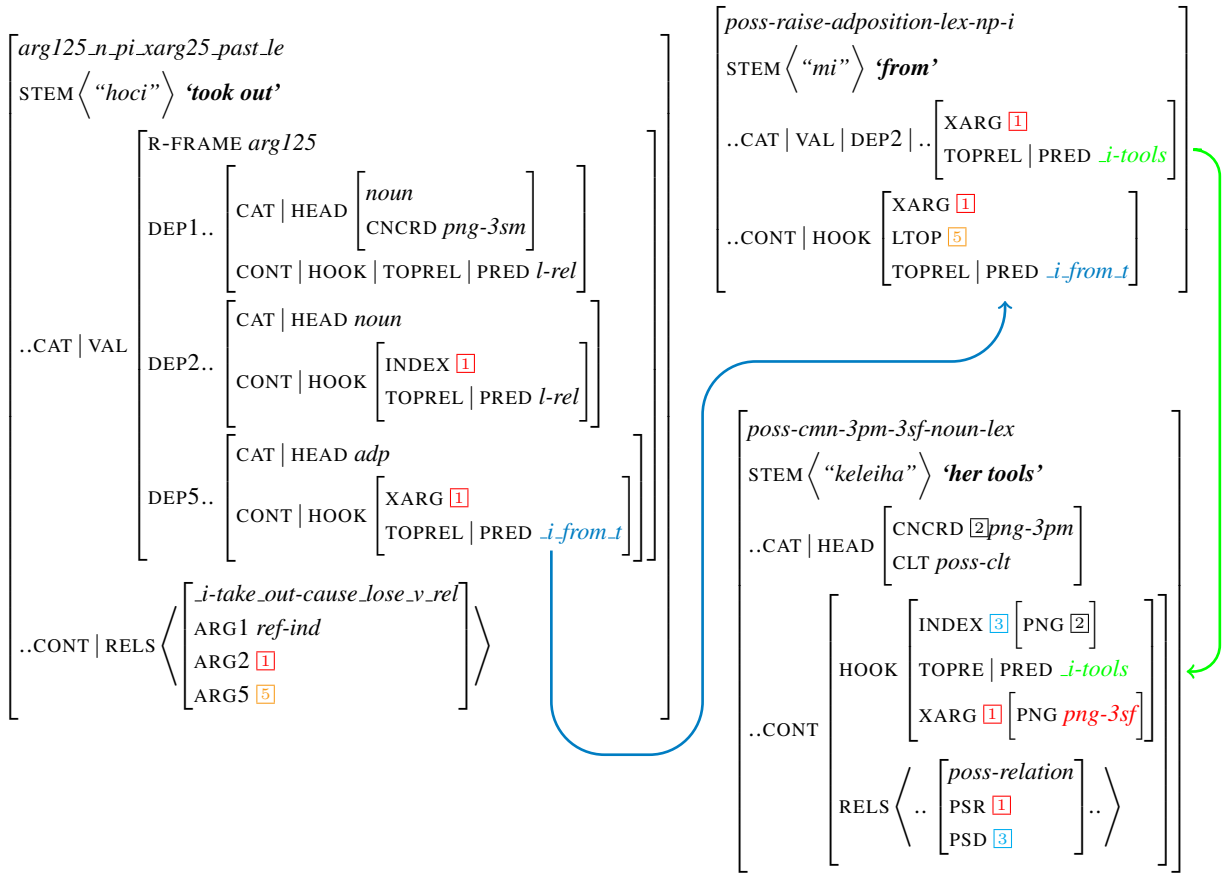

Figure 1: Co-indexed possessive idioms: The selection chain

the one used in an MWE selects for a specific complement, whereas the standard preposition does not.

[3] Bond et al. (2013) introduce an extra *identity* relation to the semantics of idiomatic verbs, which identifies the possessor and the index of the appropriate argument. This solution requires “MRS rewriting rules”, which are not needed in our analysis.

The **subcategorization** properties of *hoci* ('*take.out*') are expressed in its VALENCE, which includes the three relevant arguments: DEP1 (*Actor*), DEP2 (*Theme*) and DEP5 (*Source*) (the rest are suppressed for space reasons). Moreover, the value of its R(EALIZATION)-FRAME is *arg125*, indicating that all arguments are obligatory.[4] **Selectional restrictions** regarding the head of the PP complement are defined in the TOPREL feature of the *synsem* under DEP5. For the idiomatic *hoci* ('*take.out*') the value is the idiomatic semantic relation *_m_from_tools_ip_rel* (abbreviated *_i_from_t*) and for its literal counterpart it is the literal *_from_p_rel* (not shown). The **non-local selection** of the complement inside the idiomatic PP is made local by the lexical definition of the selecting preposition *mi* ('*from*'), which, in turn, constrains the TOPREL of its DEP2 complement to be the idiomatic *_i-tools-temper_n_rel* (abbreviated *_i-tools*).

The **co-indexation** relation between the *Theme* argument of the verb and the possessor of the "lower" NP involves a number of constraints. First, the possessed NP projects the INDEX feature of the possessor as its XARG. Second, the preposition assumes the XARG value of its NP complement, and it is raised to the PP level. Third, the co-indexation between the INDEX feature of the NP complement and the XARG of the PP complement is lexically defined for the idiomatic *hoci* ('*take.out*'), as an instance of a general lexical type *arg125_n_pi_xarg25_past_le*, which accounts for similar possessed idioms.

4 Conclusion

We presented an account of Hebrew verbal MWEs in an existing HPSG grammar. The analysis covers a multitude of MWE types, including challenging phenomena such as (possessive) co-indexation and internal modification. Moreover, the grammar now produces two analyses for most MWEs, corresponding to their idiomatic and literal readings.

MWEs are challenging because they blur the traditional distinction between the lexicon and the grammar. In our analysis, support of MWEs required minimal changes to the grammar: most crucially, the division of *rels* to either *i-rels* or *l-rels*. All other changes involve the lexicon: we make extensive use of HPSG's type hierarchies in order to state generalizations over lexical types.

The main contribution of this work is of course the extension of the coverage of HeGram to verbal MWEs. To the best of our knowledge, this is the first account of Hebrew MWEs in a linguistically-motivated grammar. Moreover, the mechanisms that we advocate are fully applicable to other languages, and can be incorporated into existing HPSG grammars with minimal effort.

In the future we intend to explore syntactic constraints on MWEs and account for their full behavior. This includes phenomena such as topicalization, wh-questions, coordination, etc.

References

Tali Arad Greshler, Livnat Herzig Sheinfux, Nurit Melnik, and Shuly Wintner. Development of maximally reusable grammars: Parallel development of Hebrew and Arabic grammars. In *HPSG15: the International conference on Head-Driven Phrase Structure Grammar*, Singapore, August 2015.

Francis Bond, Sheefa Samara Sameha, and Dan Flickinger. Making English possessed idioms our own. Paper presented at The 20th International Conference on Head-Driven Phrase Structure Grammar, Berlin, 2013.

Ann Copestake. Representing idioms. Paper presented at The 20th International Conference on HPSG, Copenhagen, 1994.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, 2002.

[4] Optionality in HeGram is expressed by disjunctive values, e.g., the R-FRAME value of the literal *hoci* ('*take.out*') is *arg12-125*, indicating that DEP5 is optional.

Livnat Herzig Sheinfux, Nurit Melnik, and Shuly Wintner. Representing argument structure in computational grammars. Submitted, 2015.

Paul Kay and Ivan A. Sag. A lexical theory of phrasal idioms. Unpublished manuscript, Stanford University, 2012.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, 1994.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavik, Iceland, 2014. ELRA. ISBN 978-2-9517408-8-4. URL `http://www.lrec-conf.org/proceedings/lrec2014/index.html`.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico, 2002.