Accounting for Hebrew verbal MWEs in HPSG

Livnat Herzig Sheinfux Nurit Melnik Shuly Wintner

Verbal MWEs in Hebrew

We present work in progress whose goal is to incorporate verbal MWEs in an HPSG grammar of Modern Hebrew (MH) that we are currently developing. MWEs in 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. MWEs come in various syntactic patterns; in this study we focus on VPs only.

MH has rich, productive morphology which is based on a root-and-template system. There are five basic verbal templates with varying degrees of semantic coherence. Each root can be realized in a number of templates. A lemma is the combination of a root and a particular template. Each verbal lemma is associated with dozens of inflected forms. Verbs inflect for number, gender, and person and exhibit full agreement with their subjects. The unmarked constituent order of clauses is SVO, yet there are cases of VS and V2, with a more restricted distribution. The order within the VP is relatively free.

MWEs in Modern Hebrew generally respect the agreement relationships between verbs and their subjects as well as agreement within the NP. There are, however, cases where these relationships are violated. In 1, an indefinite noun is modified with a definite adjective (Al-Haj et al., 2014). Conversely, some MWEs impose agreement not just between the subject and the verb but also between the verb and a constituent within the complement (2). In this particular MWE, full agreement is required between the subject, the verb, and the reflexive possessor argument within the construct-state NP complement.

(1) *ʿayin ha-raʿ*
eye the-evil
'evil eye'

(2) *dan ʿamad ʿal daʿt-o*
dan stood.3SM on mind-his.3SM
'Dan insisted'

With respect to word order, MWEs are mostly similar to standard VPs. There are, however, some expressions which impose strict word order constraints on their complements (contrast the idiomatic use in 3 with the non-idiomatic use in 4).

(3) a. *ṭaman roš-o ba-x̱ol*
hid.3SM head-his in.the-sand
'buried his head in the sand'
b. **ṭaman ba-x̱ol roš-o*
hid.3SM in.the-sand head-his

(4) a. *ṭaman mixtav ba-argaz*
hid.3SM letter in.the-box
b. *ṭaman ba-argaz mixtav*
hid.3SM in.the-box letter
'hid a letter in the box'

The relationship between MWEs and Hebrew morphology is explored by Horvath and Siloni (2009), who examine whether MWEs allow for *verbal diatheses*, i.e., alternations of a given root between different templates. They show that some MWEs allow for verbal diatheses (5) but others do not (6).

(5) a. *yaca me-ha-kelim*
went.out from-the-dishes
'got very angry'
b. *hoci et X me-ha-kelim*
took.out ACC X from-the-dishes
'made X very angry'

(6) a. *higdil roš*
made.grow head
'assumed responsibility'
b. *roš-o gadal*
head-his grew
only literal: 'His head grew bigger'

The aforementioned examples are only a small sample of the types of non-standard behaviors which MWEs in Hebrew exhibit. Additional issues include limited internal modifications, fossil words, non-standard inflection, syntactic irregularity, and more (Al-Haj et al., 2014).

Description of the current grammar

We are currently in the process of developing HeGram, a deep linguistic processing grammar of Modern Hebrew. HeGram is grounded in the theoretical framework of Head-driven Phrase Structure Grammar (HPSG, Pollard and Sag (1994)) and is implemented in the LKB system (Copestake, 1999, 2002).

The architecture of the grammar embodies significant changes to the way argument structure is standardly viewed in HPSG. Its main contribution is that 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*).

In a nutshell, this involves the distribution of valence features across ten categories. Each valence category is characterized both in terms of its semantic role, as well as the types of syntactic phrases which can realize it. Consequently, the semantic relations which predicates denote consist of coherent argument roles, which are consistent across all predicates in the language. This is a clear departure from the minimalistic semantic representation schema employed by computational HPSG grammars implemented in the framework of the DELPH-IN[1] initiative, which involves four argument labels, assigned consecutively to arguments regardless of their semantic role (Copestake, 2009).

Incorporating MWEs in the current grammar

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). Nevertheless, our motivation stems not just from pragmatic reasons. We view MWEs as a challenging test case for the innovative architecture that we have implemented in HeGram.

One phenomenon whose account is improved by our system 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. Barring cases of fossilized words, verbs which head VP MWEs 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 5b subcategorizes for two complements: *Theme* and *Source*. Syntactically, the two complements are realized as NP and PP, respectively. Besides the obvious difference between the 'standard' and the idiomatic *hoci*, which is that the latter restricts its Source argument to only one item (*me-ha-kelim* '*from the dishes*'), an additional difference is that the Source argument of the 'standard' verb is optional (e.g., *hocia et ha-zevel* '*took out the garbage*'). Consequently, the two senses are each listed separately in the lexicon as instances of two subtypes of a common supertype, each with its respective sense-specific constraints.

We expect that the finer semantic distinctions that our system allows, as well as the explicit differentiation between syntactic constraints and semantic constraints, will prove to be beneficial to the incorporation of MWEs into the grammar.

[1] http://www.delph-in.net/

References

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.

Copestake, A. (1999). The (new) LKB system. Technical report, Stanford University.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Copestake, A. (2009). Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9. Association for Computational Linguistics.

Horvath, J. and Siloni, T. (2009). Hebrew idioms: The organization of the lexical component. *Brill's annual of Afroasiatic languages and linguistics*, 1(1):283–310.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). 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.