Towards Hebrew FrameNet

Miriam R. L. Petruck

Miriam R. L. Petruck holds a Ph.D. in Linguistics from the University of California, Berkeley. She has been connected with FrameNet, at the International Computer Science Institute in Berkeley, CA, since 1998. Her current work to develop Hebrew FrameNet brings her back to her earliest interest in modern Hebrew semantics, syntax and the lexicon.
miriamp@icsi.berkeley.edu

Thanks to a publicly available Hebrew language (newspaper) corpus[1], as well as other web-based resources, such as Rav-Milim[2] and Hebrew WordNet[3], the creation of Hebrew FrameNet (HFN) has become possible. Moreover, there are good prospects for cooperation and collaboration with the computational linguistics community in Israel (e.g. access to larger corpora for research and evaluation purposes or use of software for lemmatization and search).

In the context of research that investigates the universality of the semantic frame, initial steps have been taken towards the development of HFN, an on-line lexical resource for contemporary Hebrew which will provide the semantic and syntactic combinatorial possibilites, or *valences*, for each item analyzed, through the manual annotation of example sentences in a newspaper corpus (and eventually, the automatic capture and organization of the annotation results for web-based viewing and querying). With advances in computer technology and the existence of (searchable) corpora, the work of lexicography has changed dramatically. The fine-grained semantic classification and syntagmatic information of the sort to be provided by HFN will make its database an invaluable resource for lexicographers and advanced language teachers/learners, as well as researchers in linguistics and natural language processing (NLP).

In accord with FrameNet, the first computational lexicography project of its kind[4] (Fontenelle 2003), HFN is based on the principles of Frame Semantics (FS; Fillmore 1977, 1985, Petruck 1996), at the heart of which is the *semantic frame*, an experience-based schematization of the speaker's world against word meaning can be understood. In FS, a linguistic unit evokes a frame, whose *frame elements* (FEs), or participants and props in a scene, indicate the semantic roles that need to be filled. An FS analysis of a lexical item relies on the identification and definition of the frame(s) in which the word participates, along with the frame-specific FEs that are recorded as triples of information about the semantic role, the phrase type and the grammatical function of the annotated constitutent.

Initially, HFN will produce full annotation for frame evoking elements[5] in the newspaper corpus. This serves as a means of (1) creating the infrastructure for using the FrameNet DeskTop (FNDT; a suite of tools used for defining frames, FEs, and words, and annotating example sentences) for the analysis of Hebrew texts and (2) determining the level of linguistic description and computational representation at which the lexicon of modern Hebrew can be characterized in terms of existing FS concepts. Adapting FNDT for HFN will demonstrate the feasibility of using the software for a non-IndoEuropean language.[6] Investigating the linguistic expression of events and scenarios through the same or different frames will also document the different lexicalization patterns of Hebrew and English (Talmy 2000).

As with FrameNets for other languages (e.g. Spanish[7]), the HFN database will function as both a dictionary and a thesaurus. The dictionary-like features include definitions, tables summarizing the patterns of syntactic realizations of FEs occurring with a word, and sets of annotated sentences from the corpus showing the semantic information associated with each syntactic pattern. Like a thesaurus, words are linked to the semantic frames in which they participate, and frames are linked to other collections of words and to related frames. FS analyses are useful for research in crosslinguistic lexicology (Subirats and Petruck 2003) and in the advanced foreign language classroom (Sato 2004). The availability of such information via the internet will facilitate studies in Hebrew linguistics as well as Hebrew language teaching/learning. Ultimately, HFN data will serve the needs of research in NLP for Hebrew, contributing deep semantic information for a variety of tasks, including word sense disambiguation, machine translation, information extraction and question answering (Likowski 2004).

Notes

1. http://mila.cs.technion.ac.il/website/english/resources/corpora/2000sentences/index.html
2. http://ravmilim.co.il; see *Kernerman Dictionary News* 12, 2004.
3. http://multiwordnet.itc.it/online
4. http://icsi.berkeley.edu/~framenet
5. A frame evoking element is any sense of a word that brings to mind a frame.
6. Similar efforts are under way for Japanese, see http://nak.ics.keio.ac.jp/jfn.
7. http://gemini.uab.es/sfn

The complete version of this article, including the full references as well as a practical example, is available online: http://kdictionaries.com/kdn/kdn1304.html